PORTFOLIO MANAGEMENT RESEARCH

SUMMER 2020 VOLUME 2 NUMBER 3

the journal of financial data science

JFDS.pm-research.com

expanding the frontiers of data-driven investing

—FRANK J. FABOZZI, MARCOS LÓPEZ DE PRADO, AND JOSEPH SIMONIAN

In turbulent conditions harness the power of ESG financial intelligence

THE JOURNAL OF
impact&esg
INVESTING

launching September 2020

With environmental, social, and governance factors becoming increasingly important when making investment decisions and managing risk, our new journal will ensure you have all the research you need to manage and enhance your portfolios effectively.

JESG.pm-research.com

PORTFOLIO
MANAGEMENT
RESEARCH

the journal of financial data science

VOLUME 2, NUMBER 3 SUMMER 2020

Frank J. Fabozzi	Editor
Marcos López de Prado	Editor
Joseph Simonian	Editor
Francesco A. Fabozzi	Managing Editor
Mitchell Gang	Production Editor
Deborah Brouwer	Production and Design Manager
Mark Adelson	Content Director
Rosie Instance	Marketing Manager
William Law	Account Manager
Nikol Madjarova	Account Manager
Ryan C. Meyers	Account Manager
David Rowe	Reprints Manager and Advertising Director

In the lead article in this issue, Dilip B. Madan and Yazid M. Sharaiha explain the linkages that may exist between automated trading and financial markets, both in theoretical and practical contexts. As they explain in their article, "Machine Trading: *Theory, Advances, and Applications*," dynamic contributions to trading are evaluated over some investment horizon using covariations between asset position and price changes. They compare machine learning strategies based on a Gaussian process regression with a least squares regression. The authors generalize these two regression methods by invoking conservative valuation schemes, leading to the study of conservative conditional expectations modeled by distorted expectations which, in turn, lead to the development of distorted least squares and distorted Gaussian process regression as the associated estimation or prediction schemes. The authors analyze trading strategies based on these four regression methods—Gaussian process regression, distorted Gaussian process regression, least squares regression, and distorted least squares—using a common database of securities and set of factor (predictor) variables. The regression methods are used to trigger entry into trades, with exits conducted on a comparable basis for all of them. Trading strategies are executed for nine sectors of the US economy using 14 different predictive factor sets. Results reported by the authors indicate improvements are made by Gaussian process regression over the least squares regression, and distorted least squares regression, with the distorted regression also favorably affecting the drawdowns.

Many financial models and applications rely on the ability to quickly determine which hidden state a new time-series observation belongs to. In "Greedy Online Classification of Persistent Market States Using Realized Intraday Volatility Features," Peter Nystrup, Petter N. Kolm, and Erik Lindström develop a greedy online classifier for classifying time-series data without the need to parse historical data. While it is difficult to maintain persistence when classifying observations online, their classification methodology is based on clustering temporal features and explicitly penalizing jumps between states with a fixed-cost regularization term. A series of simulations show that the new classifier results in a higher accuracy and increased robustness to misspecification than the correctly specified maximum likelihood estimator.

Hyperparameter optimization has become increasingly important as the number of machine-learning approaches to portfolio selection continues to increase. In "Hyperparameter Optimization for Portfolio Selection," Peter Nystrup, Erik Lindström, and Henrik Madsen propose a systematic approach to hyperparameter optimization by leveraging recent advances in automated machine learning and multi-objective optimization. They establish a connection between forecast uncertainty

and holding- and trading-cost parameters. Nystrup, Lindström, and Madsen argue that these should be considered regularization parameters that can be adjusted to achieve optimal performance. In particular, they show that when the signal-to-noise ratio of return forecasts deteriorates, the optimal level of the transaction-cost parameters increases to prevent excessive trading. Additionally, the authors show that hyperparameter optimization using multi-objective optimization produces solutions with better risk-return tradeoffs than that of manual, grid, and random search. At the same time, the solutions found were more stable across in-sample training and out-of-sample testing, indicating that improvements were not a result of overfitting.

There is an extensive literature that has investigated hedge funds and hedge fund interconnectedness on an index level. However, what ultimately determines fund interconnectedness is single hedge fund strategies and their underlying risk factors. Yet, few studies have examined this interconnectedness. In "A Network Approach to Analyzing Hedge Fund Connectivity," Gueorgui S. Konstantinov and Joseph Simonian investigate this issue. Specifically, they investigate the hedge fund market as a network of interacting individual funds. Identifying and analyzing the most important hedge fund styles that could both affect the market and transmit systemwide shocks to other funds, individual asset classes, and beyond, they find that the most connected hedge fund database categories are global macro and equity long–short funds. Applying clustering algorithms to classify and evaluate hedge fund styles as well as individual hedge funds, they present empirical evidence that hedge fund networks are found to have different dynamics across different stages of the business cycle. Classifying funds using clustering, in which seemingly different funds are shown to cluster based on their shared factor exposures, is a major finding of their study. Rather than database classifications, when measuring hedge fund risk across the business cycle, this finding of Konstantinov and Simonian suggests that investors should consider fund connectivity and their attendant importance scores. Moreover, they propose a forecasting framework that can be used to predict hedge fund network behavior and the impact of individual factors on the network.

Financial data distributions change over time due to short-lived market dynamics and exogenous drivers among others. This is often manifested through nonlinear patterns that are known to be challenging to track by conventional models. In "Derivation of a Dynamic Market Risk Signal Using Kernel PCA and Machine Learning," Alireza Yazdani applies Kernel PCA on the spot return covariance matrix of G10 currencies, alongside other known drivers of the foreign-exchange market, to derive an aggregate market signal that can be used as an input feature in currency prediction models. Yazdani demonstrates that the Kernel PCA signal enhances in-sample and out-of-sample risk-adjusted performance across a range of machine learning strategies. He attributes the performance to the market "information conveyance" and "information compression" properties of the Kernel PCA method as well as the ability to track short- and mid-term market moves. This is particularly useful as a predictor of extreme events when utilized in a machine learning framework.

While the predictability of financial market returns is studied extensively in the literature, many studies focus on individual return factors and linear models. Such studies tend to make three strong parametric assumptions that undermine predictive performance: (1) a simple functional form between returns and predictors, (2) no interactions within the features, and (3) the feature set usually includes less than a handful of predictors. In "The Cross Section of Commodity Returns: *A Nonparametric Approach*", Clemens Struck and Enoch Cheng compare the predictability of commodity returns using standard linear approaches with that of tree-based methods which overcome the three parametric assumptions listed above. Struck and Cheng find in their out-of-sample analysis that up to 3.74% of returns are predictable using tree-based methods versus 0.33% using the standard linear approaches. Portfolios formulated using tree-based methods outperform their linear counterparts on both an absolute and risk-adjusted basis.

Multiple machine learning models have been found to enhance predictive performance. Combining models is commonly referred to as "committees". Rather than use the traditional approach to portfolio selection, Tsung-wu Ho proposes a way that the committee approach can be applied in portfolio selection, the most

popular optimization frameworks for portfolio selection being risk-diversification, risk parity, and risk optimal. The basic methodology Ho uses in "Portfolio Selection Using Portfolio Committees," is as follows. Since each optimal portfolio is a combination of three basic elements (strategy, covariance matrix, and risk type), the combination is augmented to 250 optimal portfolios at each estimation period. Then a score as an information criterion is defined by Ho to select the best portfolio to hold in the next period. The author finds that the committee approach to portfolio selection provides superior performance. Moreover, it easy to implement.

The hierarchical risk parity approach introduced in 2016 by Marcos López de Prado elegantly addresses the weaknesses of portfolio models and establishes this approach as the cutting-edge application of machine learning to portfolio management. In "A Modified Hierarchical Risk Parity Framework for Portfolio Management," Marat Molyboga extends the original hierarchical risk parity approach by incorporating three intuitive elements commonly used by practitioners: (1) an exponentially weighted covariance matrix with Ledoit-Wolf shrinkage instead of the sample covariance matrix, (2) an equal volatility, rather than an inverse variance, allocation approach to diversify across portfolio constituents both within and across clusters, and (3) portfolio volatility targeting to diversify across time. Examining the impact of the enhancements on portfolios of Commodity Trading Advisors within a large-scale Monte-Carlo simulation framework that accounts for the realistic constraints of institutional investors, the author reports a striking improvement in the out-of-sample Sharpe ratio of 50% (on average), as well as a reduction in downside risk. Molyboga argues that the hierarchical risk parity framework as modified in the article has broad applications for portfolios of traditional and alternative investments.

Francesco A. Fabozzi
Managing Editor

the journal of financial data science

Editors

Frank J. Fabozzi
EDHEC Business School

Marcos López de Prado
True Positive Technologies
and Cornell University

Joseph Simonian
Autonomous Investment
Technologies

Managing Editor

Francesco A. Fabozzi
Frank J. Fabozzi Associates

Advisory Board

Irene Aldridge
Cornell University and AbleMarkets

Robert Arnott
Research Affiliates

Joseph Cerniglia
University of Pennsylvania

Rama Cont
Oxford University

Germán G. Creamer
Stevens Institute of Technology

Sanjiv Ranjan Das
Santa Clara University

Campbell Harvey
Fuqua School of Business
Duke University

Michael Imerman
Peter F. Drucker and Masatoshi Ito
Graduate School of Management
Claremont Graduate University

Bruce Jacobs
Jacobs Levy Equity Management

Ron Kahn
BlackRock Financial Management

Kathryn Kaminski
AlphaSimplex Group LLC

Hossein Kazemi
CISDM/Isenberg School
of Management and
University of Massachusetts at Amherst

Petter Kolm
Courant Institute,
New York University

Han Liu
Northwestern University

Anthony Ledford
Man-AHL

Yongjae Lee
Ulsan National Institute
of Science and Technology

Chen Liang
Bridgewater Associates

Jim Kyung-Soo Liew
Johns Hopkins University

Ganesh Mani
Carnegie Mellon

Dhagash Mehta
The Vanguard Group

Paul F. Mende
MIT Sloan School of Management

George Mussalli
PanAgora

Abdolreza Nazemi
Karlsruhe Institute of Technology (KIT)

Warren Pennington
Vanguard

Sidney C. Porter
FEV Analytics

Arun Verma
Bloomberg

Zhipeng (Alan) Yan
Martin Tuchman School of Management,
New Jersey Institute
of Technology

The Journal of Financial Data Science®, (ISSN# 2640-3943) is published quarterly by Pageant Media, Ltd., 41 Madison Avenue, 20th Floor, New York, NY 10010. Phone (646) 891-2110. Copyright © 2020 Pageant Media, Ltd. POSTMASTER: Send address changes for *The Journal of Financial Data Science* to Pageant Media, Ltd., 41 Madison Avenue, 20th Floor, New York, NY 10010 USA, Att: Journals Department. Reproduction in whole or in part without written permission is prohibited.

Manuscript: *The Journal of Financial Data Science* considers but assumes no responsibility for unsolicited manuscripts. Unsolicited manuscripts will not be returned. Instructions for submission of manuscripts are printed at the end of the journal.

Subscriptions: For subscription enquiries please contact a member of the team at +1 646 931 9045/+44 20 7139 1600 or at pm-research@pageantmedia.com. Printed by The Manson Group, St Albans, UK.

The opinions expressed are those of the individual authors. Portfolio Management Research is not responsible for the accuracy, completeness, or timeliness of the information contained in the articles herein. If you require expert advice, you should seek the services of a competent professional. No statement in this journal is to be construed as a recommendation to buy or sell securities. Product names mentioned in the articles in this journal may be trademarks or service marks of their respective owners.

Machine Trading: *Theory, Advances, and Applications* 8

Dilip B. Madan and Yazid M. Sharaiha

Dynamic contributions to trading are evaluated using covariations between position and price changes over a horizon. Other performance measures such as Sharpe ratios, gain–loss ratios, acceptability indexes, and drawdowns are also employed. Machine learning strategies based on Gaussian process regression (GPR) are compared with least squares (LSQ). Furthermore, both are generalized by invoking conservative valuation schemes that lead to the study of conservative conditional expectations modeled by distorted expectations. The latter lead to the development of distorted least squares (DLSQ) and distorted Gaussian process regression (DGPR) as the associated estimation or prediction schemes. Trading strategies are executed for nine sectors of the US economy using 14 different predictive factor sets. Results indicate improvements are made by GPR, DGPR over LSQ, and DLSQ, with the distorted versions also favorably affecting the drawdowns.

TOPICS: ***Statistical methods, simulations, big data/machine learning****

Greedy Online Classification of Persistent Market States Using Realized Intraday Volatility Features 25

Peter Nystrup, Petter N. Kolm, and Erik Lindström

In many financial applications, it is important to classify time-series data without any latency while maintaining persistence in the identified states. The authors propose a greedy online classifier that contemporaneously determines which hidden state a new observation belongs to without the need to parse historical observations and without compromising persistence. Their classifier is based on the idea of clustering temporal features while explicitly penalizing jumps between states by a fixed-cost regularization term that can be calibrated to achieve a desired level of persistence. Through a series of return simulations, the authors show that in most settings their new classifier remarkably obtains a higher accuracy than the correctly specified maximum likelihood estimator. They illustrate that the new classifier is more robust to misspecification and yields state sequences that are significantly more persistent both in and out of sample. They demonstrate how classification accuracy can be further improved by including features that are based on intraday data. Finally, the authors apply the new classifier to estimate persistent states of the S&P 500 Index.

TOPICS: ***Statistical methods, simulations, big data/machine learning****

Hyperparameter Optimization for Portfolio Selection 40

Peter Nystrup, Erik Lindström, and Henrik Madsen

Portfolio selection involves a trade-off between maximizing expected return and minimizing risk. In practice, useful formulations also include various costs and constraints that regularize the problem and reduce the risk due to estimation errors, resulting in solutions that depend on a number of hyperparameters. As the number of hyperparameters grows, selecting their value becomes increasingly important and difficult. In this article, the authors propose a systematic approach to hyperparameter optimization by leveraging recent advances in automated machine learning and multiobjective optimization. They optimize hyperparameters on a train set to yield the best result subject to market-determined realized costs. In applications to single- and multiperiod portfolio selection, they show that sequential hyperparameter optimization finds solutions with better risk–return trade-offs than manual, grid, and random search over hyperparameters using fewer function evaluations. At the same time, the solutions found are more stable from in-sample training to out-of-sample testing, suggesting

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

they are less likely to be extremities that randomly happened to yield good performance in training.

TOPICS: ***Portfolio theory, portfolio construction, big data/machine learning****

A Network Approach to Analyzing Hedge Fund Connectivity 55

Gueorgui S. Konstantinov and Joseph Simonian

In this article, the authors investigate the hedge fund market as a network of interacting individual funds. The authors identify and analyze the most important hedge fund styles that could both affect the market and transmit systemwide shocks to other funds, individual asset classes, and beyond. The authors find that the most connected hedge fund database categories are global macro and equity long–short funds. A central result of the article is a classification of funds using clustering, in which seemingly different funds are shown to cluster based on their shared factor exposures. This finding demonstrates that investors should consider fund connectivity and their attendant importance scores rather than database classifications when measuring hedge fund risk across the business cycle. The authors also provide a forecasting framework that can be used to predict hedge fund network behavior and the impact of individual factors on the network.

TOPICS: ***Analysis of individual factors/risk premia, factor-based models, style investing, performance measurement****

Derivation of a Dynamic Market Risk Signal Using Kernel PCA and Machine Learning 73

Alireza Yazdani

Kernel principal component analysis (PCA) is an extension of the conventional PCA method that employs a kernel transformation whereby hidden patterns in possibly multidimensional data may be detected and extracted more explicitly. In this article, the author applies the method of kernel PCA to a currency prediction case study and derives an aggregate market signal. It is observed that this signal has desirable information-compression properties and may be used as a predictive risk indicator in the return prediction models. Used alongside common drivers of exchange rates, a kernel PCA signal enhances in-sample and out-of-sample risk-adjusted performance across a range of machine learning strategies. In particular, the author observes that a kernel PCA signal remains robust and predictive during volatile market conditions. The kernel PCA signal may be used as a machine learning feature to inform and support data-driven risk management strategies.

TOPICS: ***Big data/machine learning, derivatives****

The Cross Section of Commodity Returns: *A Nonparametric Approach* 86

Clemens Struck and Enoch Cheng

To what extent are financial market returns predictable? Standard approaches to asset pricing make strong parametric assumptions that undermine their return-predicting ability. The authors employ tree-based methods to overcome these limitations and attempt to approximate an upper bound for the predictability of returns in commodities futures markets. Out of sample, they find that up to 3.74% of 1-month returns are predictable—more than a 10-fold increase from standard approaches. The findings hint at the importance multiway interactions and nonlinearities acquire in the data; they imply that new factors should be tested on their ability to add explanatory power to an ensemble of existing factors.

TOPICS: ***Futures and forward contracts, commodities****

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

Portfolio Selection Using Portfolio Committees 104

Tsung-wu Ho

The author proposes a committee approach to portfolio selection. Because each optimal portfolio is a combination of three basic elements—strategy, covariance matrix, and risk type—the author first augments the combination to 250 optimal portfolios at each estimation period. The author then defines a score to select the best portfolio to hold in the next period. Survival of the fittest, the superior performance of the combination portfolio, demonstrates that the committee approach to portfolio selection is not only effective but also easy to implement.

TOPICS: ***Portfolio theory, portfolio construction****

A Modified Hierarchical Risk Parity Framework for Portfolio Management 128

Marat Molyboga

This article introduces a modified hierarchical risk parity (MHRP) approach that extends the HRP approach by incorporating three intuitive elements commonly used by practitioners. The new approach (1) replaces the sample covariance matrix with an exponentially weighted covariance matrix with Ledoit–Wolf shrinkage; (2) improves diversification across portfolio constituents both within and across clusters by relying on an equal volatility, rather than an inverse variance, allocation approach; and (3) improves diversification across time by applying volatility targeting to portfolios. The author examines the impact of the enhancements on portfolios of commodity trading advisors within a large-scale Monte Carlo simulation framework that accounts for the realistic constraints of institutional investors. The author finds a striking improvement in the out-of-sample Sharpe ratio of 50%, on average, along with a reduction in downside risk.

TOPICS: ***Statistical methods, simulations, big data/machine learning****

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

Machine Trading: *Theory, Advances, and Applications*

Dilip B. Madan and Yazid M. Sharaiha

Dilip B. Madan
is professor emeritus at the Robert H. Smith School of Business in College Park, MD.
dbm@rhsmith.umd.edu

Yazid M. Sharaiha
is global head of Systematic Strategies at Norges Bank Investment Management in London, UK.
ysh@nbim.no

KEY FINDINGS

- The authors demonstrate how finance theory reduced the potential contributions of data science by assumption.
- The concept of conservative conditional expectations is defined, developed, and implemented to enhance predictive technologies such as least squares and Gaussian process regression.
- Using factor attributes as input, the trading of stock return predictions classically and conservatively illustrates the practical benefits of the advocated theoretical advances.

ABSTRACT: *Dynamic contributions to trading are evaluated using covariations between position and price changes over a horizon. Other performance measures such as Sharpe ratios, gain–loss ratios, acceptability indexes, and drawdowns are also employed. Machine learning strategies based on Gaussian process regression (GPR) are compared with least squares (LSQ). Furthermore, both are generalized by invoking conservative valuation schemes that lead to the study of conservative conditional expectations modeled by distorted expectations. The latter lead to the development of distorted least squares (DLSQ) and distorted Gaussian process regression (DGPR) as the associated estimation or prediction schemes. Trading strategies are executed for nine sectors of the US economy using 14 different predictive factor sets. Results indicate improvements are made by GPR, DGPR over LSQ, and DLSQ, with the distorted versions also favorably affecting the drawdowns.*

TOPICS: ***Statistical methods, simulations, big data/machine learning****

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

Machine learning methodologies have developed powerful techniques enabling machines to learn or synthesize numerous complex functional relationships between multidimensional inputs and outputs, when in fact such relationships exist. The inputs could be marks on a two-dimensional grid of high dimension, and the output, an indicator of which of the 26 letters of the alphabet has been written. Every person can read the letter that has been written, so there must be a clear functional relationship between the inputs and the output.

Machines can now be taught to read the alphabet correctly. Beyond reading the alphabet, machines can be taught to play complex games such as chess and go and beat human players. In modern automation, machines are being taught to perform many human tasks, with robots replacing humans in driving cars and flying planes. A natural

question that then arises is whether machines can be taught to be superior to humans as traders of financial assets. They can certainly process decisions on many more assets at much greater speeds, but will their activities only yield losses? Many attempts have been made to build successful automated traders, and some definitely have succeeded even if many have failed. Our objective here is to better understand and explain the linkages that may be present between automated trading and financial markets, both in their theoretical and practical contexts.

Broadly speaking, the output of interest is trade profitability, and the inputs are variables used to define asset positions through time. Critical to the exercise is comprehension of the functional relationship, if any, that exists between the inputs and the output if a machine is to be asked to learn such a function. In dynamic markets, the relevant inputs and functional relationships may be changing, with an associated need for continuous relearning. The first step is then an analysis of the relationship between trade profitability and asset positioning. The classical formulation of this relationship in the finance literature is that the gains from trading are a stochastic integral of positions against the gains in asset values. Specifically, they exclude terms of covariation between changes in positions and gains in asset values. The latter is where data science methods may seek to make their main contributions to trade profitability; the assumptions of modern financial analysis set this contribution to zero. Our first contribution recognizes the relevance of such covariation contributions.

There will then be two separate contributions of positioning to profitability. The first, classical, contribution is related to positioning with positive alpha. The second is related to building positive covariance between changes in positions and asset values. In developing and exploring the second contribution, we apply classical machine learning methods of Gaussian process regression (GPR) to link positioning to selected driving variables. Alternatively, one could also use support vector machines and neural nets, but in the interest of parsimony and for the purpose of illustrating the issues, we restrict our attention to just GPR. However, the least squares (LSQ) objective of GPR treats positive and negative errors symmetrically, and this is not appropriate when dealing with financial matters that translate errors to financial gains and losses. Aversion to losses and preference for gains lead us to modify the concept of LSQ to distorted LSQ (DLSQ), which differentiates the treatment of losses and gains. The DLSQ objective results in constructing a distorted GPR (DGPR). The second contribution of this article is the introduction of DGPR.

Trading strategies may thus be based on LSQ, DLSQ, GPR, and DGPR. All four are implemented on a common database of securities to be traded using a common set of predictor or factor variables. The methods are used to trigger entry into trades, with exits conducted on a comparable basis for all of them. Exits are based on a maximum holding period curtailed by reaching profit targets or stop loss levels. This article concludes with a comparative analysis of trading performance achieved by the four prediction systems. There are 85 sets of factor variables employed for all securities traded in 14 different subsets of 5 to 12 variables. These are proprietary variables, the details of which cannot be revealed.

The outline of the rest of the article is as follows. The second section addresses the relationship between trade profitability and asset positioning. The third section introduces GPR and its generalization to DGPR via DLSQ. The fourth section presents the details for the implementation of trading strategies based on these predictors. The fifth section presents the results for trading based on the four training criteria employed. The sixth section concludes.

PROFITABILITY AND POSITIONING

Consider the position n_i shares in asset i with market value S_i. The current value of all holdings, V, is

$$V = \sum_i n_i S_i$$

After some time, there will be a change in positions and market values to positions n_i' with values S_i' and V', the market value of all holdings now being

$$V' = \sum_i n_i' S_i'$$

For notational convenience, define the inner product $\langle x, y \rangle$, of vectors $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots y_n)$ as

$$\langle x, y \rangle = \sum_i x_i y_i$$

We may then write the profitability or gain/loss in value as

$$\begin{aligned} V' - V &= \langle n', S' \rangle - \langle n, S \rangle \\ &= \langle n, S' - S \rangle + \langle S, n' - n \rangle + \langle n' - n, S' - S \rangle \\ &= B + C + D \end{aligned}$$

There are then three components to profitability, identified as B, C, and D. The first term, B, is the buy-and-hold component of profitability were the original positions to be held constant in the interim. Note that this is the case when n' would have to equal n. The second term, C, is the cost of switching positions immediately to n' at the prices. The third term, D, is the benefit of dynamic trading, or acquiring assets that increased in value and selling those that decreased in value.

Data science methods are designed to enhance value and contribute to profitability by enhancing D. For example, one may use a prediction model to predict that asset prices will take the final values $\hat{S}$. The predictions may then be employed to define position changes by

$$n' - n = g(\hat{S} - S)$$

for some monotone increasing function g that is zero at zero. The resulting contribution to the change in value or profitability is the dynamic contribution δ that we may define as

$$\delta = E\left\langle g(\hat{S}' - S), (S' - S) \right\rangle \tag{1}$$

Good active trading strategies are those with high dynamic δ contributions. Good static gains reflect initial positions n that are long the assets that increase in value and short those that decrease, over the interim. The classical high alpha positions are those for which

$$\begin{aligned} \alpha &= E[B] \\ &= E(\langle n, S' - S \rangle) \end{aligned}$$

is high.

With a view toward reviving focus on dynamic trading contributions, possibly to be enhanced by the methods of data science, we briefly survey how they were lost or eliminated from financial analysis in the first place. This absence occurred in the literature on discrete time models, followed by developments in continuous time models. An appreciation of the theoretical status of data science methods in financial analysis is critical to the development of these methods and their financial applications.

Data Science in Financial Mathematics: Discrete Time

In the discrete time context, numerous authors and textbooks on financial mathematics amalgamate terms C and D to equivalently write

$$V' - V = \langle n, S' - S \rangle + \langle S', n' - n \rangle$$

Examples include Cvitanić and Zapatero (2004), Epps (2007), Skiadas (2009), and Jarrow (2018). The view is then taken that the new positions are adopted at the end of the interval at the prices S'. Self-financing portfolios are then defined as those for which

$$\langle S', n' - n \rangle = 0$$

For such self-financing strategies

$$\begin{aligned} V' - V &= \langle n, S' - S \rangle \\ &= B \end{aligned}$$

is the buy-and-hold gain.

Somewhat differently, Shiryaev (1999) amalgamated one B term and three D terms to write

$$V' - V = \langle n', S' - S \rangle + \langle S, n' - n \rangle$$

Now, if new positions are taken at the prices S, self-financing is then defined by the condition

$$\langle S, n' - n \rangle = 0$$

with the gain from trade being

$$B + D = \langle n', S' - S \rangle$$

and the expected profitability being $\alpha + \delta$.

However, in actuality, new positions are taken neither at the start of the period nor at the end but instead at numerous time points in the interim at the

prices prevailing at these times. However, one cannot refer to these prices in a discrete time model. Thus, let us consider interim time points as well. These considerations take us to the literature on continuous time models.

Data Science in Financial Mathematics: Continuous Time

The next steps may be taken in the context of continuous time representations of these results. First, with the convention of C and D amalgamated into the self-financing condition, the continuous time representation is clear and we have that

$$dV = \langle n, dS \rangle \tag{2}$$

with all gains/losses being of the buy-and-hold form.

Alternatively, if B and D are amalgamated, in continuous terms we would have the representation

$$dV = \langle n, dS \rangle + \langle dn, dS \rangle$$

Now suppose that price processes are semi-martingales (Jacod and Shiryaev 2000; Protter 2005) that may be decomposed into their predictable and martingale components with $S = A + M$, where A is predictable and M is a martingale. In this case, the instantaneous dynamic component is

$$\langle dn, dS \rangle = \langle dn, dA + dM \rangle$$

If n and A are predictable and absolutely continuous with respect to the Lebesgue measure, then both $\langle dn, dA \rangle$ and $\langle dn, dM \rangle$ are zero—the first being of order $(dt)^2$ and the second by virtue of no covariations being possible between predictable and martingale components. We are thus back to the buy-and-hold gains (Equation 2) for the expected instantaneous profitability.

Although the dynamic component is lost in the instantaneous representations, it can reappear in the analysis of results related to the aggregation of time over horizons of interest. For the analysis of such time aggregations, we shall consider both positions and prices as being pure jump processes possibly adapted to common drivers.

Data Science and Profitability: Aggregated Continuous Time

It is well recognized that trade profitability requires the passage of time. In very short time spans, there is only a loss incurred in paying the spreads and other transaction costs. Over time, the process for positions $n(t)$ may be seen as a pure jump process with

$$n(t) = \sum_{s \le t} \Delta n(s)$$

Here $\Delta n(s)$ is the change in positions undertaken at time s. In general, one could consider as a candidate a local structure that is a compound Poisson process with adapted arrival rates $\lambda(\omega)$ for the occurrence of a trade with an adapted distribution $g(\omega, x)$ for the number of shares to be traded. The process $n(t)$ then has compensator

$$v_n(\omega, x)dxdt = \lambda(\omega)g(\omega, x)dxdt$$

The process $n(t)$ is here seen as a pure jump process with increments occurring at unpredictable times. Classically, $n(t)$ is modeled as a predictable process with no martingale component and hence orthogonal to all martingale increments. Many trades, however, are initiated by machines scouring news sources for items on companies, with executions beginning within seconds of news signals. Given that trades may be triggered by news events occurring at possibly unpredictable times, with subsequent execution at a high speed, we allow the paths of positions to have compensated jump martingale components.

Additionally, in keeping with the modeling presented by Madan and Schoutens (2016), we take the price process to be a pure jump process with

$$S(t) = \int_0^t S(u_) \int_{-\infty}^{\infty} (e^y - 1)\mu(dy, du)$$

where $\mu(dy, du)$ is the integer valued random measure for the jumps in the logarithm of the stock price. We suppose that μ has the compensator $\nu(\omega, dy, du)$. There is an extensive literature on pure jump finite variation stock price models. Examples include the normal inverse Gaussian model (Barndorff-Nielsen 1998), the variance gamma model (Madan and Seneta 1990; Madan, Carr, and Chang 1998), the CGMY model (Carr et al. 2002),

and the generalized hyperbolic model (Eberlein and Keller 1995; Eberlein 2001). For more recent examples, we refer to Küchler and Tappe (2008); Madan and Wang (2017); and Madan, Schoutens, and Wang (2018).

It is reasonable to suppose that the two jump processes for n and S have no common or simultaneous jumps. The latter would reflect pure arbitrages. The two processes then have zero covariation. For the dynamic benefit, we wish to evaluate for δ the expression

$$\delta = E\left[\langle n(h) - n, S(h) - S\rangle\right]$$

Covariation and Covariance

The analysis here follows that presented by Madan and Schoutens (2018) with adjustments for context. For two processes $Y = (Y(t),\ t \geq 0)$ for prices and $n = (n(t),\ t \geq 0)$ for positions, both pure jump processes, the joint arrival rate function for the pair (Y, n) in the absence of covariation takes the special form

$$k_{Y,n}(\omega, t, y, x) = k_Y(\omega, t, y)\mathbf{1}_{x=0} + k_n(\omega, t, x)\mathbf{1}_{y=0}$$

where the dependence on ω is adapted.

We suppose the sums of all jumps in both the Y and n processes are well defined, as is the case for finite variation processes. The joint characteristic function at horizon h is given by

$$\begin{aligned}
&\phi_{Y(h),n(h)}(u,v) \\
&\quad = E[\exp(iuY(h) + ivn(h))] \\
&\quad = E\left[\exp\left(\int_0^h ds \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (e^{iuy+ivx} - 1)k_{Y,n}(\omega, s, y, x)dydx\right)\right]
\end{aligned}$$

From the absence of covariation, it follows that

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (e^{iuy+ivx} - 1)k_{Y,n}(\omega, t, y, x)dydx \\
&= \int_{-\infty}^{\infty} (e^{iuy} - 1)k_Y(\omega, t, y)dy + \int_{-\infty}^{\infty} (e^{ivx} - 1)k_n(\omega, t, x)dx
\end{aligned}$$

The covariance of log returns with position innovations for horizon h is then given by

$$\begin{aligned}
&-\frac{\partial^2}{\partial u \partial v}\phi_{Y(h),n(h)}(u,v)|_{u=v=0} \\
&+\left(\frac{\partial}{\partial u}\phi_{Y(h),n(h)}(u,v)|_{u=v=0}\right)\left(\frac{\partial}{\partial v}\phi_{Y(h),n(h)}(u,v)|_{u=v=0}\right)
\end{aligned}$$

Evaluating the required derivatives yields the results

$$\frac{\partial}{\partial u}\phi_{Y(h),n(h)}(u,v)|_{u=v=0} = iE\left[\int_0^h ds \int_{-\infty}^{\infty} yk_Y(\omega, s, y)dy\right]$$

$$\frac{\partial}{\partial v}\phi_{Y(h),n(h)}(u,v)|_{u=v=0} = iE\left[\int_0^h ds \int_{-\infty}^{\infty} xk_n(\omega, s, x)dx\right]$$

Define the stochastic instantaneous variations by

$$v_Y(\omega, s) = \int_{-\infty}^{\infty} yk_Y(\omega, s, y)dy$$

$$v_n(\omega, s) = \int_{-\infty}^{\infty} xk_n(\omega, s, x)dx$$

Furthermore, evaluate

$$\begin{aligned}
&-\frac{\partial^2}{\partial u \partial v}\phi_{Y(h),n(h)}(u,v)|_{u=v=0} \\
&\quad = E\left[\int_0^h ds \int_{-\infty}^{\infty} yk_Y(\omega, s, y)dy \int_0^h ds \int_{-\infty}^{\infty} xk_n(\omega, s, x)dx\right] \\
&\quad = E\left[\int_0^h dsv_Y(\omega, s) \int_0^h dsv_n(\omega, s)\right]
\end{aligned}$$

The covariance is thus given by

$$\begin{aligned}
Cov(Y(h), n(h)) &= E\left[\int_0^h dsv_Y(\omega, s) \int_0^h dsv_n(\omega, s)\right] \\
&\quad - E\left[\int_0^h dsv_Y(\omega, s)\right] E\left[\int_0^h dsv_n(\omega, s)\right]
\end{aligned}$$

Define the integrated instantaneous variations by

$$V_Y(\omega, t) = \int_0^t dsv_Y(\omega, s)$$

$$V_n(\omega, t) = \int_0^t dsv_n(\omega, s)$$

We then observe that

$$E\left[\langle n(h) - n, Y(h)\rangle\right] = E\left[V_n(\omega, h)V_Y(\omega, h)\right] \tag{3}$$

or that position price covariations are the expectation of the product of the integrated variations. If the variations are stochastic, this exceeds the product of the expected variations by the covariance between these integrated variations.

The dynamic profitability of an invested dollar is, however, given by

$$\delta = E\left[\langle n(h) - n, \exp(Y(h)) - 1\rangle\right]$$

and Equation 3 is a first-order approximation to the profitability. The objectives of data science–based trading strategies are to enhance profitability via the δ for the trade and go beyond α. The order of the cross variation of Equation 3 may be that of h^2, and it could be argued that this is to be ignored as h tends to zero. Realistically, however, h remains significant because, even if one is working with 1-minute data with holding periods of 10 minutes, the unit time is then a minute and h is 10. Covariations at short horizons may arise through the stochasticity of variations, even though in the instantaneous time limit they are zero. We later explore these time aggregation issues from a numerical perspective.

Illustration of Dynamic Contributions

We wish to exhibit covariation-based dynamic contribution to wealth accumulation. For this purpose, we take two independent square root processes (Cox, Ingersoll, and Ross 1985) driven by independent Brownian motions as signal processes:

$$dx_t^1 = \kappa_1(\theta_1 - x_t^1)dt + \sigma_1\sqrt{x_1}dW_t^1$$
$$dx_t^2 = \kappa_2(\theta_2 - x_t^2)dt + \sigma_2\sqrt{x_2}dW_t^2$$

We take the process of share accumulation to be pure jump with unit jumps and arrival rate $a_1 + b_1 x_t^1$. The process for share decumulation is also pure jump with unit jumps and arrival rate $a_2 + b_2 x_t^2$. Let the process for the number of shares held be n_t, where

$$n_t = n_{1t} - n_{2t}$$

is the difference of processes with unit jumps.

The local motion of the logarithm of the stock price is taken to lie in the bilateral gamma process class (Küchler and Tappe 2008; Madan and Wang 2017; Madan, Schoutens, and Wang 2017, 2020). This is a four-parameter process written as the difference of two independent gamma processes, each of which describes the process for the positive and negative price moves, respectively. The scale and speed of the positive moves are denoted by the parameters b_p, c_p, whereas for the negative moves the parameters are b_n, c_n, respectively. When the up and down speeds are made linear in state variables x_t^1, x_t^2, respectively, the local drift is then

$$b_p c_p(x_t^1) - b_n c_n(x_t^2).$$

We suppose that

$$c_p(x_t^1) = \alpha_1 + \beta_1 x_t^1$$
$$c_n(x_t^2) = \alpha_2 + \beta_2 x_t^2$$

For the square root process

$$dx = \kappa(\theta - x)dt + \lambda\sqrt{x}dW$$

with $4\kappa\theta > \lambda^2$ we use the scheme (Andersen, Jäckel, and Kahl 2010)

$$x_{t+\Delta} = \frac{x_t + \kappa\theta\Delta + \lambda\sqrt{x_t}Z\sqrt{\Delta} + \frac{1}{4}\lambda^2\Delta(Z^2 - 1)}{1 + \kappa\Delta}$$

Given paths of x_t^1, x_t^2, we may generate n_{1t}, n_{2t}, X_t and evaluate the covariance of Δn, ΔS.

For the setting

$$a_1 = 200$$
$$b_1 = 250$$
$$a_2 = 200$$
$$b_2 = 150$$
$$\alpha_1 = 1$$
$$\beta_1 = 0.75$$
$$\alpha_2 = 1$$
$$\beta_2 = 0.25$$
$$\theta_1 = 1$$
$$\theta_2 = 1$$
$$\kappa_1 = 0.5$$
$$\kappa_2 = 0.75$$
$$\lambda_1 = 0.2$$
$$\lambda_2 = 0.3$$

The time step was $h = 1/252$. The correlation between Δn and ΔS across 200 days was 0.4356.

We simulated 10,000 paths with a length of 2,500 days for the signal processes x_{1t}, x_{2t}. For interval lengths

of h days for h ranging from two days to one year, we evaluated the static contribution to the change in wealth as follows. Let $n_{i,t}$ be the number of shares held on path i at time t. Let $p_{i,t}$ be the corresponding price. The static contribution $sc_{i,t}$ is defined by

$$sc_{i,t} = n_{i,t}(p_{i,t+h} - p_{i,t})$$

For the dynamic contribution we define

$$dc_{i,t} = (n_{i,t+h} - n_{i,t})(p_{i,t+h} - p_{i,t})$$

The cost of trade is defined by

$$c_{i,t} = p_{i,t}(n_{i,t+h} - n_{i,t})$$

and the change in wealth is

$$\Delta w_{i,t} = n_{i,t+h}p_{i,t+h} - n_{i,t}p_{i,t}$$

The self-financed change in wealth is

$$\Delta\tilde{w}_{i,t} = \Delta w_{i,t} - c_{i,t}$$

We have identity

$$\Delta\tilde{w}_{i,t} = sc_{i,t} + dc_{i,t}$$

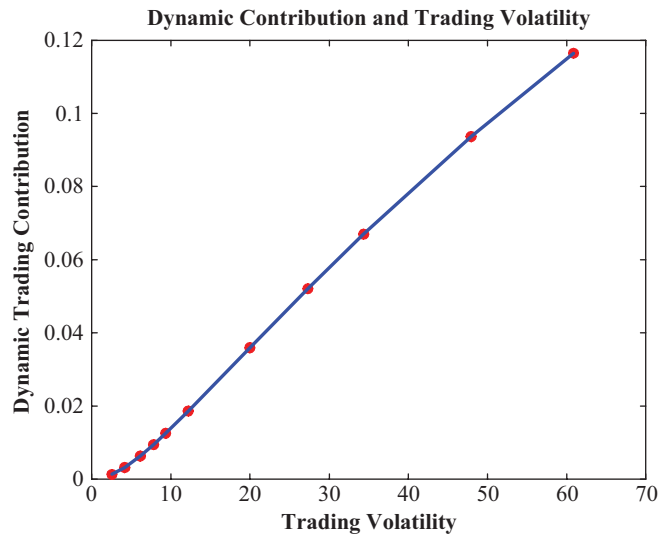
The dynamic proportion is defined by the average over time of the average across paths of the dynamic contribution to the average self-financed change in wealth. The trading volatility is defined by the average over time of the standard deviation across paths of $n_{i,t+h} - n_{it}$. Exhibit 1 presents a graph of the dynamic contribution against the trading volatility. The trading volatility does go to zero as the time step is decreased, and the dynamic contribution is zero in the limit. For realistic time steps encountered in data analysis, however, the dynamic contribution may be significant.

PREDICTION TECHNOLOGIES

Data science and machine learning methodologies have introduced numerous new powerful technologies for functional representation. One of the potential functions of interest in financial trading contexts is the expectation of financial returns conditional on current market information. They generalize the use of regression using linear functions of input data to that

EXHIBIT 1

Graph of the Dynamic Contribution to Change in Wealth As a Function of the Trading Volatility



of linear functions of so-called *nonlinear feature transforms* of input data. The number and form of feature transforms that may be employed are quite wide and varied. Three popular technologies are GPR (Rasmussen and Williams 2006), support vector machines (SVM) (Hastie, Tibshirani, and Friedman 2009), and neural nets (NN) (Bishop 1995). Each of these technologies aims at teaching a machine to learn a functional relationship between inputs and outputs using different ways of constructing features from the input data and possibly different ways of evaluating the errors associated with the model predictions. Applications on datasets often then report on the prediction quality associated with different input sets, choices of functional representation, and training criteria used. From the perspective of financial trading applications, however, the prediction per se is not the financial objective. It has to be monetized into a trading strategy, and the interest is in the profitability of the latter. We report on the performance of such monetized strategies; this is closer to the ultimate objective, which is, of course, the actual performance in practice.

Of the three technologies, GPR, SVM, and NN, we shall focus on just GPR. Before proceeding, we make some comparative remarks. On large datasets, GPR can become expensive in computation times; the computation cost is related to the need to invert

large matrixes. SVM, by employing methods of linear programming, avoids these issues. However, SVM introduces other design issues, such as the choice of error sizes to be ignored and regularization parameters to be used. Preliminary comparisons in our applications favored GPR over SVM. Furthermore, for the sample sizes we examine, which use a half year or one year of prior daily data on a rolling basis, the GPR computation costs are not an issue. Additionally, we note that GPR and SVM do use similar feature constructions.

With regard to NN, it may be observed that the feature constructions are quite different. NN functions are general and can, with sufficient parameters, synthesize even Dirac delta functions. They can then go much further in the direction of overfitting data, whereas the GPR functions are based on Gaussian kernels that have been used widely for functional interpolation, smoothing, and extrapolation. At this stage of our investigation, we thought it best to steer away from overfitting issues and to work with the slightly more constrained feature construction associated with GPR.

GPR Review

We present a brief review of GPR and refer the reader to Rasmussen and Williams (2006) for further details. Suppose we wish to train a machine to learn a real-valued function of n variables, $y = f(x)$. For the training, we suppose we have available values y_i that have been observed for the function with errors at the points x_i, $i = 1, \ldots, N$. Let (X, y) be the $N \times n$ matrix with rows x_i and $y = (y_1, \ldots, y_N)$ is the column vector of the observed y values. The underlying error hypothesis is that

$$y_i = f(x_i) + \varepsilon_i$$

where ε_i are independently and identically distributed normal random variables with zero mean and variance σ^2. The variance parameter is estimated by maximum likelihood.

We now come to the Gaussian process assumptions on $f(x)$. These assumptions are on the vectors $v = (x, f(x))$. It is supposed that for any finite set of vectors $v_j = (x_j, f(x_j))$, $j = 1, \ldots, J$ the J-dimensional vector $(f(x_1), \ldots, f(x_J))$ is multivariate and normally distributed with mean $(m(x_1), \ldots m(x_J))$ and covariance matrix

$$K_{j,j'} = k(x_j, x_{j'}),\ 1 \le j, j' \le J$$

for a prespecified covariance kernel function $k(x, x')$. One often works with a mean function that is either identically zero or just a vector of constant means for each component.

There are many choices possible for the kernel function, and a popular choice is a Gaussian kernel function that in one dimension takes the form

$$k(x, x') = \sigma_f^2 \exp\left(-\sum_{i=1}^{n} \frac{(x_i - x_i')^2}{2h_i} \right)$$

where $h = (h_1, \ldots, h_n)$ is a vector of bandwidth parameters. The parameters σ_f^2, h are hyperparameters that may be estimated by maximum likelihood.

Basically, the further away x is from x', the lower the covariance between the outputs $f(x)$ and $f(x')$. Thus, one may learn about the function at a particular x_* by observing the values seen in its neighborhood.

For our dataset (X, y) and a zero mean vector, we first note that $\mathbf{f} = (f(x_1), \ldots f(x_N))$ is multivariate normal with zero mean and covariance matrix

$$K(X, X) = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ k(x_2, x_1) & \cdots & k(x_2, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix}$$

For other points x_{*m}, $m = 1, \ldots, M$ forming the $M \times n$ matrix X_* and the vector $\mathbf{f}_* = f(x_{*1}, \ldots, x_{*M})$, the joint distribution of $(y, \mathbf{f}_*)$ is multivariate normal with mean zero and covariance matrix

$$\begin{pmatrix} K(X, X) + \sigma^2 I_N & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{pmatrix},$$

where I_N is the N-dimensional identity matrix.

The conditional expectation of $\mathbf{f}_*$ given X_*, X, y is

$$E\left[\mathbf{f}_* | X_*, X, y)\right] = K(X_*, X)\left[K(X, X) + \sigma^2 I_N\right]^{-1} y$$

with conditional covariance matrix

$$K\left(X_*, X_*\right) - K(X_*, X)\left[K(X, X) + \sigma^2 I_N\right]^{-1} K(X, X_*).$$

The hyperparameters σ_f^2, h are to be estimated by maximizing the marginal log likelihood given by

$$-\frac{1}{2}\log(\det(K(X,X)) - \frac{1}{2}\gamma' K(X,X)^{-1}\gamma$$

The mean functions may be revised from zero or a constant using basis functions $h(x)$ with parameters β to

$$m(x) = h(x)'\beta$$

A typical set of basis functions is the linear and quadratic functions in the arguments x.

Prediction to Positioning

Consider the conditional expectation of a future asset price

$$Y(t) = E^P[S(t+h)|\mathcal{F}_t]$$

If at time t we hold the position

$$n(t) = Y(t) - S(t)$$

then on unwind at time $t + h$ the profit is

$$\begin{aligned} &n(t)(S(t+h) - S(t)) \\ &= (Y(t) - S(t))(S(t+h) - S(t)) \end{aligned}$$

and the expected profit from such positioning is

$$(Y(t) - S(t))^2 > 0.$$

The actual profit, however, is a random variable

$$X = (Y(t) - S(t))(S(t+h) - S(t))$$

and we may be interested not in its statistical expected value but its market value. The latter is given by expectations under the risk-neutral or pricing measure Q or $E^Q[S(t+h)|\mathcal{F}_t]$. In efficient markets, one has

$$S(t) = E^Q[S(t+h) \mid \mathcal{F}_t]$$

and hence X is identically zero for any position. One may have a gap temporarily between $S(t)$ and $E^Q[S(t+h)|\mathcal{F}_t]$, with the difference reflecting a positive or negative alpha situation. For such trades, one has to adopt some asset pricing model, compute α, and then base one's trades on these entities. Even if one did this, however, it may be unwise to fix attention on a single asset pricing model; temporally relevant risk factors and premiums can vary stochastically, and what looks like positive alpha on a single measure may in fact be negative alpha on the correct model for the moment.

Conservative Prediction and Positioning

An alternative prediction technology is to employ the valuation theories of two price economies as set out by Madan and Schoutens (2016) and the references therein. In this theory, the use of conditional expectations as the prediction object is replaced by the use of nonlinear conditional expectations. Two nonlinear expectations are developed as candidates for a lower and an upper prediction. Basically, if the lower expected return is positive, then the support for a long position is stronger. Similarly, when the upper expected return is negative, the argument for taking a short position is enhanced.

The lower nonlinear expectation $Y_L(t)$ is defined by

$$\begin{aligned} Y_L(t) &= \mathcal{E}^P\left[S(t+h) \mid \mathcal{F}_t\right] \\ &= \inf_{Q \in \mathcal{M}} E^Q\left[S(t+h) \mid \mathcal{F}_t\right] \end{aligned}$$

for a set of test or scenario measures $\mathcal{M}$. Similarly, one defines the upper nonlinear expectation $Y_U(t)$ as

$$\begin{aligned} Y_U(t) &= \sup_{Q \in \mathcal{M}} E^Q\left[S(t+h) \mid \mathcal{F}_t\right] \\ &= -\mathcal{E}^P\left[-S(t+h) \mid \mathcal{F}_t\right] \end{aligned}$$

The last equality follows from the relationship between the suprema and infima of expectations over sets of candidates. Computation of Y_L, Y_U requires an operational specification of the set of test measures $\mathcal{M}$, and we later introduce such sets.

With regard to positioning, a long position may be taken if $Y_L(t) > S(t)$ is of a size proportional to the difference to access the unwind cash flow of

$$(Y_L(t) - S(t))(S(t+h) - S(t))$$

that has a lower valuation of $(Y_L(t) - S(t))^2 > 0$.

Equivalently, if $Y_U(t) < S(t)$, a short position of $(S(t) - Y_U(t))$, accesses the unwind cash flow

$$(S(t)-Y_U(t))(S(t)-S(t+h))$$

the value of which is

$$(S(t)-Y_U(t))(S(t)-(-\mathcal{E}^P[-S(t+h)|\mathcal{F}_t]))$$
$$(S(t)-Y_U(t))^2>0$$

there is no trade to be undertaken when

$$Y_L(t)<S(t)<Y_U(t)$$

Distorted Expectations

Operational forms of the lower and upper valuations arise on employing distorted expectations. For this purpose, let $\Psi(u)$ be a concave distribution function defined on the unit interval. A cash flow X with distribution function $F_X(x)$ has a distorted distribution function given by $\Psi(F_X(x))$. Expectations under this distortion are defined by

$$\mathcal{D}^P(X)=\int_{-\infty}^{\infty} x d\Psi(F_X(x))$$

and for Ψ equal to the identity function, there is no distortion; the distorted expectation is just the expectation. However, for a proper concave distortion, one observes on integration by parts that

$$\mathcal{D}^P(X)=-\int_{-\infty}^{0}\Psi(F_X(x))dx+\int_0^{\infty}(1-\Psi(F_X(x)))dx \qquad (4)$$

For losses, the distorted distribution is raised, whereas for gains it is lowered; hence the distorted expectation is always below the expectation. Equation 4 may be written more symmetrically on defining

$$\hat{F}_X(x)=1-F_X(x)$$
$$\hat{\Psi}(u)=1-\Psi(1-u)$$

We then have that

$$\mathcal{D}^P(X)=-\int_{-\infty}^{0}\Psi(F_X(x))dx+\int_0^{\infty}\hat{\Psi}\left(\hat{F}_X(x)\right)dx$$

The function $\hat{\Psi}$ is convex and below the identity, whereas Ψ is concave and above the identity—hence the lift on losses and the drop on gains.

Distorted expectations are examples of conservative valuations with the set of test measures being all measures Q equivalent to P that satisfy for all sets A that

$$Q(A)\le\Psi(P(A))$$

In fact, Kusuoka (2001) showed that all lower valuations that depend on just the distribution function and are additive for co-monotone risks must be a distorted expectation for some distortion. Two risks X, Y are co-monotone if they have no negative co-movements. For such risks, co-monotone additivity is

$$\mathcal{E}^P(X+Y)=\mathcal{E}^P(X)+\mathcal{E}^P(Y) \qquad (5)$$

In general, the left-hand side of Equation 5 exceeds the right-hand side.

The lower valuation may be constructed using distorted expectations by

$$Y_L(t)=\mathcal{D}^P[S(t+h)|\mathcal{F}_t]$$

For the upper valuation one may define

$$\mathcal{U}^P(X)=-\int_{-\infty}^{0}\hat{\Psi}(F_X(x))dx+\int_0^{\infty}\Psi\left(\hat{F}_X(x)\right)dx,$$
$$Y_U(t)=\mathcal{U}^P\left[S(t+h)\mid\mathcal{F}_t\right]$$

The distortion function we shall employ is called *minmaxvar* and was introduced by Cherny and Madan (2009). It is defined by

$$\Psi^{(\gamma)}(u)=1-(1-u^{\frac{1}{1+\gamma}})^{1+\gamma} \qquad (6)$$

The function has an infinite derivative at zero and a zero derivative at unity, thereby reflecting a strong aversion to large losses and a heavy discounting of extremely large gains. The stress parameter γ controls the degree of concavity being embedded into the distortions.

DLSQ

Conditional expectations are estimated by the least squares criterion on recognizing that the minimum of the LSQ objective $E[(y-a)^2]$ with respect to a occurs at

$$a=E[y]$$

Similarly, when estimating the α quantile instead of the expected value, one may employ the objective

$$(1-\alpha)E[(a-y)^+]+\alpha E[(y-a)^+]$$

as the minimum with respect to a, which occurs at a such that

$$(1-\alpha)\Pr(y<a)-\alpha\Pr(y>a)=0$$

or that

$$\Pr(y<a)=\alpha$$

Coming now to distorted expectations, consider the minimization of

$$\int_{-\infty}^{\infty}(y-a)^2 d\Psi(F(y))$$

The minimum with respect to a now occurs at

$$a=\int_{-\infty}^{\infty} y d\Psi(F(y))$$

or the distorted expectation. These observations lead to the formulation of DLSQ as follows.

Consider the model

$$y=f(x,\theta)+\varepsilon_i$$

and suppose we are interested in estimating θ. Let $g(x)$ be the unknown true conditional expectation of y given x. Now observe that

$$\begin{aligned}E\left[(y-f(x,\theta))^2\right]&=E\left[(y-g(x)+g(x)-f(x,\theta))^2\right]\\&=E\left[(y-g(x))^2\right]+E\left[(g(x)-f(x,\theta))^2\right]\end{aligned}$$

and hence the least squares minimal choice of θ essentially minimizes the distance to the true unknown conditional expectation function $g(x)$.

Now let $\mathcal{D}$ be a distorted expectation operator and let

$$\mathfrak{g}(x)=\mathcal{D}(y|x)$$

and then write that

$$\mathcal{D}\left[(y-f(x,\theta))^2\right]=\mathcal{D}\left[(y-\mathfrak{g}(x))^2\right]+\mathcal{D}\left[(\mathfrak{g}(x)-f(x,\theta))^2\right]$$

so that selecting θ to minimize the distorted expectation of the squared residuals minimizes the distance to the conditional distorted expectation.

On a dataset, we proceed as follows. We minimize over θ the objective

$$\sum_i (y_i-f(x_i,\theta))^2 w_i$$

where the weights are now defined by

$$\begin{aligned}w_i&=\Psi(q_i)-\Psi\left(q_i-\frac{1}{N}\right)\\q_i&=\frac{1}{N}\sum_{j=1}^{N}\mathbf{1}_{r_j\le r_i}\\r_i&=y_i-f(x_i,\theta)\end{aligned}$$

For the upper distorted expectation, we merely replace Ψ with the convex dual distortion $\hat{\Psi}$.

DGPR

For the formulation of DGPR, we proceed as follows (*DGPR*). First we estimate a standard *GPR* that produces kernel parameters and a β mean level. Once the kernel has been estimated, the prediction for the function at the point x^* is given by

$$\hat{f}(x^*)=\beta+K(x^*,X)\alpha$$

where for *GPR* we have that

$$\alpha=(K(X,X)+\sigma^2 I)^{-1}(Y-\beta)$$

For distorted predictions, we merely alter β and α and keep intact the hyperparameters estimated by the *GPR* to construct the basis function $K(x^*,X)$. A generalization delivering distorted hyperparameters along the lines of Boukouvalas, Barillec, and Cornford (2012) could be attempted in future research.

More precisely, we solve for b, a to minimize

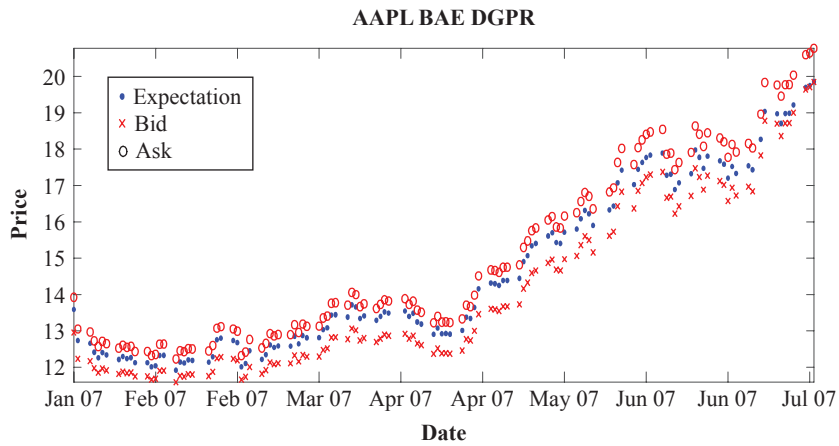
$$\sum_{i=1}^{N}(Y_i-b-K(x_i,X)a)^2\omega_i$$

where defining

$$u_i=Y_i-b-K(x_i,X)a$$

EXHIBIT 2

Bid, Ask, and Expected Values from GPR and DGPR on AAPL Prices Regressed on the Prior 10 Prices



we set

$$\omega_i = \Psi\left(q(u_i), \gamma\right) - \Psi\left(q(u_i) - \frac{1}{N}, \gamma\right)$$

with

$$q(u_i) = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{u_j \le u_i}.$$

For the upper price, the distortion Ψ is replaced by $\hat{\Psi}$, where

$$\hat{\Psi} = 1 - \Psi(1-u)$$

Application to Data on AAPL Prices

By way of a simple illustration of GPR and DGPR, we take time-series data on the prices of Apple (AAPL) from January to July of 2007 and perform a GPR and DGPR of the current price on the prior 10 prices for a sample of 126 days. Exhibit 2 presents in-sample predictions for the lower or bid, upper or ask, and expected values for AAPL. The distortion used was *minmaxvar* at a stress level of 1.5.

DESIGN OF TRADING STRATEGIES

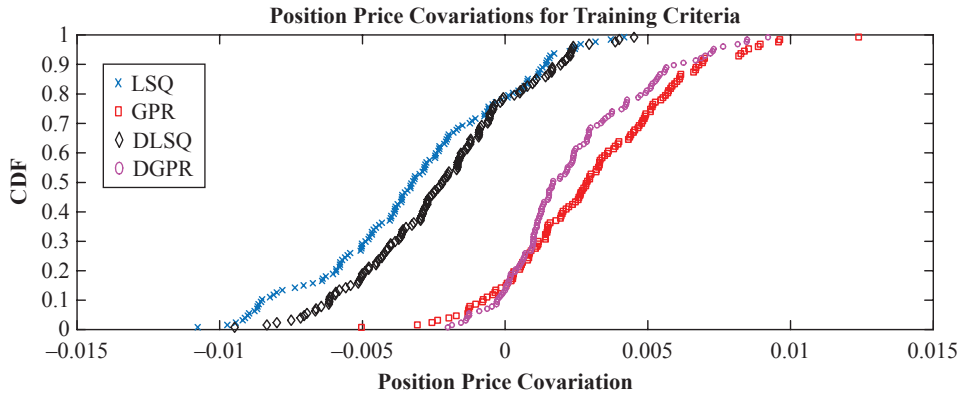
Generally, data on the performance of trading strategies report the changes in values over time, with no information at any level of detail on the positions taken. However, when implementing specific backtests of strategies, we do have position information simultaneously generated. The exercise conducted here will specifically evaluate position and price covariation or the dynamic δ contribution defined in Equation 1. In addition, we report on the more usual performance measures associated with evaluating trading strategies.

The strategies to be implemented and reported here form long–short balanced portfolios of stocks in sectors of the US economy. The portfolios entered into reverse all the positions simultaneously, and both the entry and exit from trade are accomplished at a portfolio level. There are, in all, nine sectors to be traded. The sectors are represented by a set of large midcap securities. The number of stocks potentially traded in each of the nine sectors is 14, 10, 9, 20, 21, 15, 25, 12, and 16. The stocks are ranked each day on the basis of the predicted returns, with long positions in the top-ranked stocks and short positions in the lower-ranked stocks. The position entered into is organized to be dollar neutral. The exit from positions is based on a portfolio-level setting of profit targets and stop losses with a maximum holding period of 10 days. When the profit target or stop-loss level for the portfolio is reached, all positions are reversed, and the trade is unwound.

In developing forecasts for individual stocks, the prediction inputs always include the last five immediately prior returns on the stock. In addition, some 73 prediction factors are employed in the forecasts. The prediction factors used are subsets of these 73 factors. Fourteen subsets are used for each of the nine sectors; hence, 126

EXHIBIT 3

Distribution of Covariations for Five-Day Changes in Positions with That of Prices



(9 × 14) strategies are implemented over approximately four years ending March 1, 2019. For each of the 126 sets of sector and factor set pairs, four prediction criteria of *LSQ*, *DLSQ*, *GPR,* and *DGPR* are reported. Thus, we have a total of 504 (126 × 4) trading histories for an approximate 1,000 days to be reported on.

A brief description of the factors and subsets is as follows. The factors are in six broad categories that span fundamental valuation and technical metrics. The categories are value, momentum, quality, earnings, analyst sentiment, and short interest. Each category is represented by a number of metrics, and each metric is cross-sectionally normalized on a sector-neutral basis. The first set of factors uses just a single category (e.g., using all five valuation metrics). These are book to price, free cash flow yield, sales to enterprise value, total shareholder yield, and the 12-month forward earnings yield. The latter factor sets combine factors from numerous categories. The number of factors in the 14 factor sets is, respectively, 6, 9, 6, 5, 2, 3, 3, 11, 8, 14, 13, 18, 19, and 20.

The predictors are continuously retrained on each trading date using 126 days of immediately prior input data. The training objective is to predict the immediately forward five-day return. The trained predictors are used five days later to make predictions using inputs that are not in the training sample. The profit target is set at 2% with a stop loss at 1% and a maximum holding period of 10 days. The long positions are taken in the top 30% of predictions, with short positions in the bottom 30% of predictions. No positions are taken for 40% of predictions in the interior. When working with distorted expectation predictors, we go long the top 30% of the lower predictions and short the bottom 30% of the upper predictions and remove any that are both long and short. The profit target and stop loss for balanced portfolios are based on a fraction of gross market value applied to the sum of the absolute value of all dollar positions. This was set at 0.5.

RESULTS FOR THE FOUR TRAINING CRITERIA

For each training criterion, we have a history of 993 days on the implementation of 126 trading strategies for nine sectors of the economy, with predictions based on 14 factor sets. We construct distribution functions for various performance measures across the 126 runs for each training criterion. The performance measures are as follows. First, we consider the average across stocks in the strategy for the five-day covariation between changes in positions and prices, noting that all positions are unwound in 10 days. The other criteria are at the portfolio level: the Sharpe ratio, gain–loss ratio, proportion of positive outcomes, acceptability index based on *minmaxvar* (Cherny and Madan 2009), and maximum drawdown. The Sharpe ratio is among the more widely used measures, but as noted by Cherny and Madan (2009), it is not arbitrage consistent. The gain–loss ratio corrects for this but can be tolerant of larger losses (Cherny and Madan 2009). The proportion of positive outcomes and the maximum drawdown are supplementary measures that may be employed to eliminate strategies. The acceptability index is associated with coherent risk measures as described by Artzner et al. (1999).

Exhibit 3 presents the distribution function for the average five-day covariation between changes in positions and prices. We observe that GPR and DGPR make a significant improvement over LSQ and DLSQ

Exhibit 4

Distribution of Sharpe Ratios across 126 Strategies and Four Training Criteria

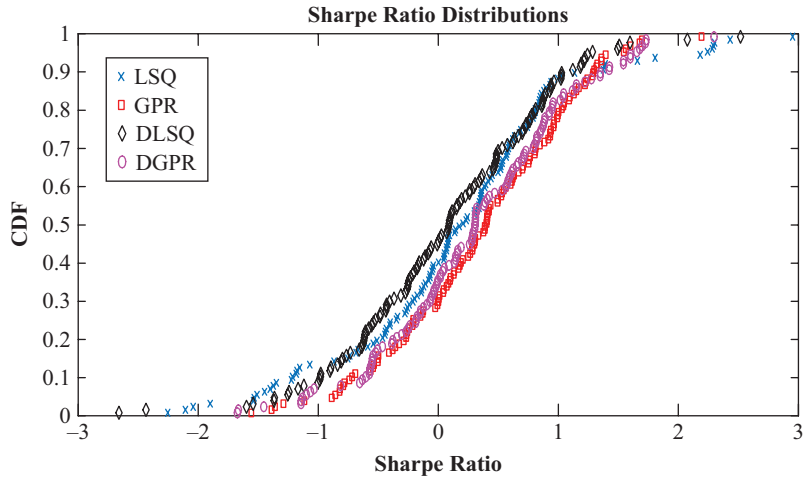
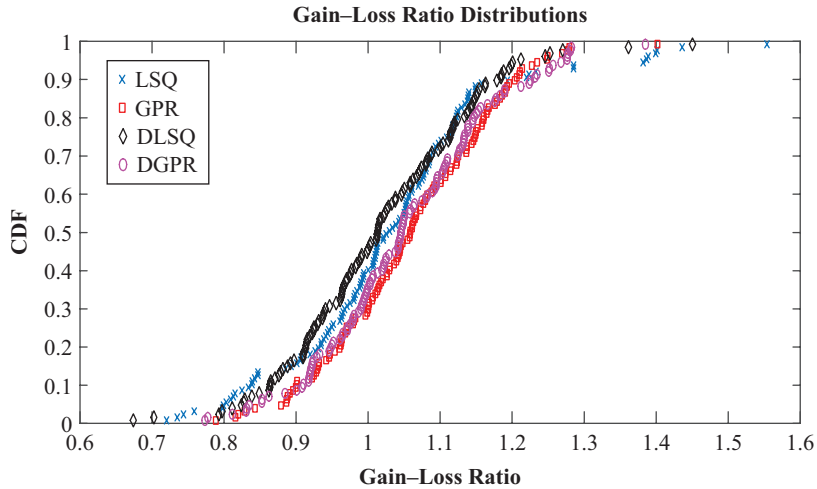


Exhibit 5

Distribution of Gain–Loss Ratio across 126 Strategies and Four Training Criteria



in this regard. Additionally, DLSQ improves on LSQ with respect to position price covariation.

Exhibit 4 displays the distribution functions for the Sharpe ratios attained. Here both DGPR and DLSQ are somewhat inferior to GPR and LSQ, with GPR dominating LSQ while DGPR dominates DLSQ.

Exhibit 5 presents the distribution function for the gain–loss ratio. The results match those for the Sharpe ratio.

Exhibit 6 presents the distribution function for the positive proportion. GPR makes an improvement over LSQ, as does DGPR relative to DLSQ, both of which are inferior to their undistorted counterparts.

Exhibit 7 presents the result for the acceptability index. Here DGPR and GPR are comparable, whereas DLSQ is inferior to LSQ. Both LSQ are dominated by Gaussian process methods.

Exhibit 8 presents the maximum drawdown over the period for the four training criteria. Here we observe that DLSQ and DGPR are comparable, with both making significant improvements over their undistorted counterparts. Additionally, GPR improves upon LSQ.

EXHIBIT 6

Distribution of the Positive Proportion across 126 Strategies and Four Training Criteria

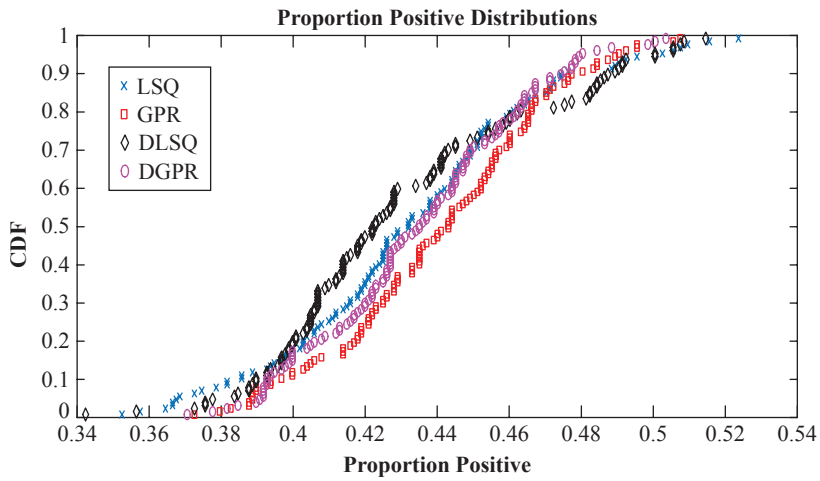
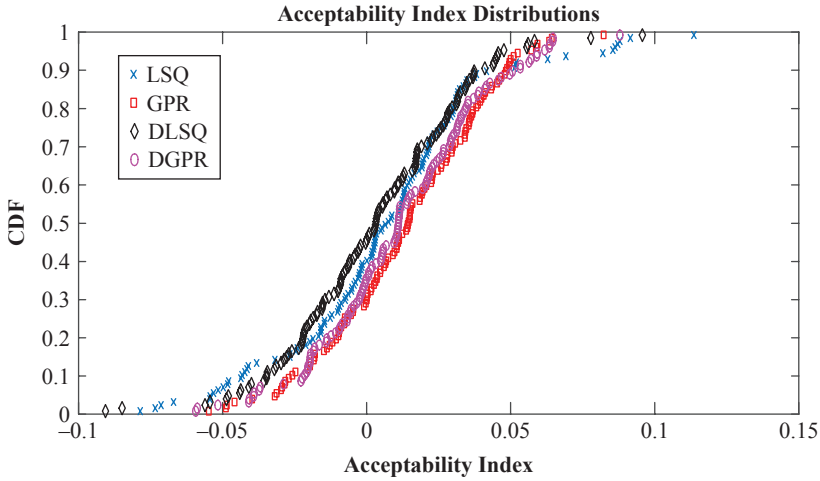


EXHIBIT 7

Distribution of Acceptability Indexes Based on *minmaxvar* Distortions across 126 Strategies and Four Training Criteria



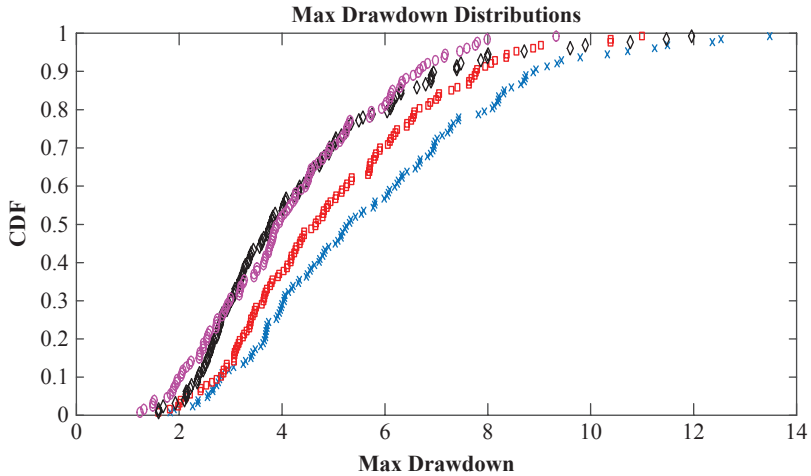
CONCLUSION

We advocate that dynamic contributions to trading be evaluated via covariations, over a horizon of interest, between changes in positions and prices. Additionally, one may consider standard measures such as Sharpe ratios, gain–loss ratios, acceptability indexes, and drawdowns. Machine learning strategies based on GPR are compared with LSQ. Furthermore, both are generalized by invoking conservative valuation schemes that lead to the study of conservative conditional expectations modeled by distorted expectations. The latter lead to the development of DLSQ and DGPR as the associated estimation or prediction schemes.

Trading strategies are executed for 9 sectors of the US economy using 14 different predictive factor sets. In each case, training is conducted on a rolling basis by LSQ, GPR, and their distorted counterparts. Results support improvements made by GPR, DGPR over LSQ, DLSQ, with the distorted versions also favorably affecting the drawdowns.

EXHIBIT 8

Distribution of Max Drawdowns across 126 Strategies and Four Training Criteria



REFERENCES

Andersen, L. B. G., P. Jäckel, and C. Kahl. "Simulation of Square-Root Processes." In *Encyclopedia of Quantitative Finance*, edited by R. Cont. Chichester, England: John Wiley and Sons. 2010.

Artzner, P., F. Delbaen, M. Eber, and D. Heath. 1999. "Coherent Measures of Risk." *Mathematical Finance* 9: 203–228.

Barndorff-Nielsen, O. E. 1998. "Processes of Normal Inverse Gaussian Type." *Finance and Stochastics* 2: 21–68.

Bishop, C. *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press. 1995.

Boukouvalas, A., R. Barillec, and D. Cornford. 2012. "Direct Gaussian Process Quantile Regression Using Expectation Propagation." Proceedings of the 29th International Conference on Machine Learning. Edinburgh, Scotland.

Carr, P., H. Geman, D. Madan, and M. Yor. 2002. "The Fine Structure of Asset Returns: An Empirical Investigation." *The Journal of Business* 75: 305–332.

Cherny, A., and D. B. Madan. 2009. "New Measures for Performance Evaluation." *The Review of Financial Studies* 22: 2571–2606.

Cox, J. C., J. E. Ingersoll, and S. A. Ross. 1985. "A Theory of the Term Structure of Interest Rates." *Econometrica* 53: 385–407.

Cvitanić, J., and F. Zapatero. *Introduction to the Economics and Mathematics of Financial Markets*. Cambridge, MA: MIT Press. 2004.

Eberlein, E. "Application of Generalized Hyperbolic Lévy Motions to Finance." In *Lévy Processes: Theory and Applications*, edited by O. E. Barndorff-Nielsen, T. Mikosch, and S. Resnick. Basel, Switzerland: Birkhäuser Verlag. 2001.

Eberlein, E., and U. Keller. 1995. "Hyperbolic Distributions in Finance." *Bernoulli* 1: 281–299.

Epps, W. *Pricing Derivative Securities*. Singapore: World Scientific. 2007.

Hastie, T., R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed. Springer Series in Statistics. Berlin: Springer. 2009.

Jacod, J., and A. Shiryaev. *Limit Theorems for Stochastic Processes*. Berlin: Springer. 2000.

Jarrow, R. A. *Continuous-Time Asset Pricing Theory*. Berlin: Springer. 2018.

Küchler, U., and S. Tappe. 2008. "Bilateral Gamma Distributions and Processes in Financial Mathematics." *Stochastic Processes and Their Applications* 118: 261–283.

Kusuoka, S. 2001. "On Law Invariant Coherent Risk Measures." *Advances in Mathematical Economics* 3: 83–95.

Madan, D. B., and W. Schoutens. 2016. *Applied Conic Finance.* Cambridge, UK: Cambridge University Press.

——. 2018. "Zero Covariation Returns." *Probability, Uncertainty and Quantitative Risk* 3: 1–29.

Madan, D., and E. Seneta. 1990. "The Variance Gamma (VG) Model for Share Market Returns." *The Journal of Business* 63: 511–524.

Madan, D. B., and K. Wang. 2017. "Asymmetries in Financial Returns." *International Journal of Financial Engineering* 4 (4).

Madan, D. B., W. Schoutens, and K. Wang. 2017. "Measuring and Monitoring the Efficiency of Markets." *International Journal and Theoretical and Applied Finance* 20 (8).

——. 2020. "Bilateral Multiple Gamma Returns: Their Risks and Rewards." *International Journal of Financial Engineering* 7 (1).

Madan, D., P. Carr, and E. Chang. 1998. "The Variance Gamma Process and Option Pricing." *The Review of Finance* 2: 79–105.

Protter, P. *Stochastic Integration and Differential Equations.* Berlin: Springer. 2005.

Rasmussen, C. E., and C. K. I. Williams. *Gaussian Processes for Machine Learning.* Cambridge, MA: MIT Press. 2006.

Shiryaev, A. *Essentials of Stochastic Finance: Facts, Models, Theory.* 3rd ed. Singapore: World Scientific Press. 1999.

Skiadas, C. *Asset Pricing Theory.* Princeton, NJ: Princeton University Press. 2009.

Disclaimer

The views expressed in this article are those of the authors and do not necessarily reflect those of Norges Bank Investment Management.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Conic Option Pricing

Dilip B. Madan and Wim Schoutens

The Journal of Derivatives
https://jod.pm-research.com/content/25/1/10

ABSTRACT: *Derivatives pricing models based on arbitrage produce a single no-arbitrage option value. In theory, trading by arbitrageurs will force the market price to equal this value. However, the real world is full of frictions and uncertainty, and there is always a range in which the market price may wander without becoming sufficiently mispriced to touch off arbitrage trades. Under weaker assumptions, as explored in this article, such a range should encompass a convex set of prices such that the net payout has positive expected value under a broad set of supporting probability distributions and is closed under scaling, which makes it a mathematical cone. Madan and Schoutens use this principle to develop a theory of* conic option pricing, *in which an upper and lower bound on an option's price are obtained by applying two deformations to the probability density for returns. Where the market price is expected to fall within these bounds is based on the principle that order flow for buying and selling should balance out on average. The resulting valuation does not require a specific density, such as lognormal (this article uses the variance-gamma process). Rather, it specifies how the density chosen by the user should be modified to produce the bounds. An empirical demonstration of the approach using seven years of options data on 208 underlying assets shows that market prices obey the pricing bounds quite well. An interesting and valuable extension is the valuation of delta and delta-gamma hedged positions, which naturally produce much tighter bounds.*

European Investing and EMU

Yazid M. Sharaiha and Michael S. Urias

The Journal of Investing
https://joi.pm-research.com/content/9/1/21

ABSTRACT: *This article examines the evidence on market integration and its implications for pan-European asset allocation by looking at evidence of declining country barriers, and the relative merits of pan-European industries as asset classes. At the same time, investors are reconsidering the merits of international diversification and global equity allocation. The authors' findings indicate that European industries are becoming less correlated than countries, on average, and offer as much, if not more, opportunity in terms of incremental performance.*

Greedy Online Classification of Persistent Market States Using Realized Intraday Volatility Features

Peter Nystrup, Petter N. Kolm, and Erik Lindström

Peter Nystrup is a postdoctoral fellow in the Centre for Mathematical Sciences at Lund University in Lund, Sweden, and in the Department of Applied Mathematics and Computer Science at the Technical University of Denmark in Lyngby, Denmark.
peter.nystrup@matstat.lu.se

Petter N. Kolm is a clinical professor and director of the Mathematics in Finance Master's Program in the Courant Institute of Mathematical Sciences at New York University in New York, NY.
petter.kolm@nyu.edu

Erik Lindström is a professor and head of the Section for Mathematical Statistics in the Centre for Mathematical Sciences at Lund University in Lund, Sweden.
erik.lindstrom@matstat.lu.se

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

KEY FINDINGS

- A new greedy online classifier is proposed that contemporaneously determines which hidden state a new observation belongs to without the need to parse historical observations and without compromising temporal persistence.
- A series of simulations demonstrates that the new classifier frequently obtains a higher accuracy and is more robust to misspecification than the correctly specified maximum likelihood estimator.
- Classification accuracy can be improved by including features that are based on intraday volatility data.

ABSTRACT: *In many financial applications, it is important to classify time-series data without any latency while maintaining persistence in the identified states. The authors propose a greedy online classifier that contemporaneously determines which hidden state a new observation belongs to without the need to parse historical observations and without compromising persistence. Their classifier is based on the idea of clustering temporal features while explicitly penalizing jumps between states by a fixed-cost regularization term that can be calibrated to achieve a desired level of persistence. Through a series of return simulations, the authors show that in most settings their new classifier remarkably obtains a higher accuracy than the correctly specified maximum likelihood estimator. They illustrate that the new classifier is more robust to misspecification and yields state sequences that are significantly more persistent both in and out of sample. They demonstrate how classification accuracy can be further improved by including features that are based on intraday data. Finally, the authors apply the new classifier to estimate persistent states of the S&P 500 Index.*

TOPICS: ***Statistical methods, simulations, big data/machine learning****

The hidden Markov model (HMM) is popular in many areas, with applications ranging from facial recognition (Nefian and Hayes 1998) and natural language processing (Gales and Young 2008; Kang, Ahn, and Lee 2018) to wind- and solar-power forecasting (Pinson et al. 2008; Bhardwaj et al. 2013), fraud detection (Robinson and Aria 2018), and risk and return modeling in finance and energy markets (Dias and Ramos 2014; Petropoulos, Chatzis, and Xanthopoulos 2016; Nystrup, Madsen, and Lindström 2018).

In many financial applications, it is highly desirable that the state sequence have a certain level of persistence. For example, if the state sequence is used as the basis of an asset allocation or trading strategy, then lower persistence leads to higher turnover and transaction costs and, consequently, poorer returns (Nystrup et al. 2015). If the model is misspecified (e.g., because the Markov assumption or the conditional distributions are incorrect) or not properly estimated (e.g., because of limited data, high dimensionality, nonstationarity, or failure to converge to the global optimum in the estimation process), typically it leads to unstable and impersistent estimates of the underlying state sequence and the introduction of unnecessary states (Bulla 2011; Fox et al. 2011; Fiecas et al. 2017). In particular, when the state sequence contains far too many jumps, the model reduces to a mixture model with no temporal information, which is of limited use in practice.

In this article, we propose a greedy online classifier that contemporaneously determines which hidden state a new observation belongs to without the need to parse historical observations and without compromising persistence. Online classification is critical for applications such as asset allocation, patient monitoring, intrusion detection, and credit card fraud monitoring, where delay after the onset of a suspicious activity might be detrimental (Nystrup et al. 2017; Garg and Pichkhadze 2019). When classifying observations online without looking at later observations, it is considerably harder to obtain persistent state estimates because of the inability to apply smoothing.

Our greedy classifier is based on the idea of clustering temporal features while explicitly penalizing jumps between states by a fixed-cost regularization term that can be calibrated to achieve a desired level of persistence. The starting point is a persistent jump model estimated using the framework proposed by Bemporad et al. (2018). Our jump model estimator builds on the work of Zheng, Li, and Xu (2019) and Nystrup, Lindström, and Madsen (2020b), incorporating a new feature set for learning HMMs that is entirely backward looking and therefore can be applied in a truly online fashion.

Through a series of simulations, we show that in most settings our new jump estimator remarkably obtains a higher accuracy than the correctly specified maximum likelihood estimator, when using the same initial parameters for the estimate. At the same time, the inferred state sequences are significantly more persistent both in and out of sample. Moreover, we demonstrate how the greedy classifier's accuracy can be further improved by including features that are based on intraday data. Finally, we compare two versions of our jump estimator to an HMM on the S&P 500 Index, where the differences in persistence are striking. Although the HMM detects state changes as quickly as the jump model, it comes at the cost of many spurious and short-lived changes. Consequently, if these estimated states were used as part of a regime-based asset allocation strategy, the lower persistence would lead to higher turnover and transaction costs.

RELATED WORK

Time-Series Segmentation

Given a sequence of time-series data, it is often desirable to partition it into states or segments, wherein each segment can be explained by as simple a model as possible. Variants of the problem have been studied in several contexts, including change-point detection (Nystrup et al. 2016), segmentation (Hallac, Nystrup, and Boyd 2019), trend filtering (Kim et al. 2009), and mixture models (Verbeek, Vlassis, and Kröse 2003; Samé et al. 2011). The task of segmenting a time series is different from clustering in general because the temporal ordering of the data must be taken into account.

Model Extensions

The classical, two-state Gaussian HMM has been extended in numerous ways to overcome its shortcomings in specific applications. Through conditional distributions with heavier tails (Bulla 2011) or time-varying parameters (Nystrup, Madsen, and Lindström 2017), it is possible to increase persistence and improve the model's ability to reproduce long memory. Other studies have relaxed the Markov assumption on sojourn-time distributions (Bulla and Bulla 2006; Langrock and Zucchini 2011), considered higher-order Markov chains (Shamshad et al. 2005), and increased the size of the hidden state space (Ghahramani and Jordan 1997; Nystrup, Madsen, and Lindström 2015). In Bayesian settings, the level of persistence can be adjusted through a hyperparameter in the hierarchical Dirichlet prior distribution imposed on the transition probability matrix (Beal, Ghahramani, and Rasmussen 2002; Fox et al. 2011).

Estimation

Standard approaches to estimating HMMs include direct numerical maximization of the likelihood function, the expectation–maximization (EM) algorithm (Baum et al. 1970; Dempster, Laird, and Rubin 1977), and Bayesian estimation using Markov chain Monte Carlo methods (Rydén 2008). The maximum likelihood estimator has favorable properties as the number of observations tends to infinity. In particular, it is consistent and asymptotically efficient; however, computationally it is not as tractable (Andersson, Rydén, and Johansson 2003).

Spectral Methods

Rousseeuw et al. (2015) used the spectral clustering algorithm of Ng, Jordan, and Weiss (2002) to estimate the parameters of a discrete-output HMM. They were able to recover the hidden partition without prior knowledge of the specific model parameters. Subsequently, Zheng, Li, and Xu (2019) used spectral clustering to learn a continuous-output HMM based on simple features constructed from the observed time series. Their results showed that spectral clustering can be more robust and more accurate than maximum likelihood estimation when the persistence in the state process is low. Nevertheless, both methods have difficulties with the highly persistent processes that are characteristic of many financial time series (Rydén, Teräsvirta, and Åsbrink 1998; Cont 2001; Nystrup, Madsen, and Lindström 2017).

Jump Penalization

Nystrup, Lindström, and Madsen (2020b) proposed a framework for fitting persistent jump models by combining the explicit penalization approach of Bemporad et al. (2018) with the temporal features for learning HMMs proposed by Zheng, Li, and Xu (2019). The jump penalty regularizes the estimation problem and provides control over the transition rate. Unfortunately, the forward-looking nature of some of the features in this approach make them less useful in practical applications where the future is unknown and observations are processed in an online fashion as they arrive.

Online State Classification

In online applications, it is hard to capture the true persistence of the state process without compromising classification accuracy. Narasimhan, Viola, and Shilman (2006) proposed an online version of the Viterbi algorithm in which a state is not classified until classification can be done with a certain confidence. When latency is permitted, it is easier to ensure persistence. Their algorithm makes the strong assumption that uncertainty in any state label decreases with latency, which is not always the case. More recently, Garg and Pichkhadze (2019) proposed an algorithm for online decoding of Markov chain models under hard latency constraints. Meanwhile, for many practical purposes it is critical to decode states online without any latency, which is the problem we consider.

JUMP MODELS

Bemporad et al. (2018) proposed to fit jump models with K states by minimizing the objective function

$$\sum_{t=1}^{T-1}[\ell(\gamma_t,\theta_{s_t})+\lambda\mathbb{I}_{s_t\neq s_{t+1}}]+\ell(\gamma_T,\theta_{s_T}), \tag{1}$$

where $\gamma = \{\gamma_1, ..., \gamma_T\}$ is a time series of T observations, the loss function $\ell(\gamma,\theta) := \|\gamma-\theta\|_2^2$ and $\mathbb{I}$ is the indicator function over the model parameters $\theta = \{\theta_1, \ldots, \theta_K\}$ and the state sequence $s = \{s_1, \ldots, s_T\}$. Intuitively, this objective function describes the trade-off between fitting the data and prior assumptions about the persistence of the state sequence.

Jump Penalty

The regularization parameter $\lambda \geq 0$ is a fixed-cost penalty for jumps. For a large enough λ, the jump model results in a single state model. When $\lambda = 0$, the model reduces to automatically splitting the dataset in at most K states and fitting one model per state, thereby generalizing the K-means algorithm (Lloyd 1982). Hence, the jump model allows us to infer the number of states from the data. This is in contrast to a standard HMM, in which the number of states is fixed a priori and model complexity is determined by rerunning the estimation for different values of K.

The jump penalty represents our prior knowledge about the state transition rate. It regularizes the objective function in Equation 1 by increasing the difference between states. In practice the jump penalty needs to be selected based on the specific application in mind—for example, by cross-validation (Nystrup, Lindström, and Madsen 2020a).

Features

Exhibit 1 outlines the construction of the features we propose for learning HMMs. In addition to the time-series observations themselves, the features consist of the last two absolute changes and rolling means and standard deviations calculated over different windows. By computing the features for two different window lengths, $l = 6$ and $l = 14$, we obtain a total of 15 features.

The features and window lengths are closely related to the features considered by Zheng, Li, and Xu (2019) and Nystrup, Lindström, and Madsen (2020b), with the important difference that they are all backward looking. The advantage of mixing forward- and backward-looking features is the inherent smoothing provided. However, in many practical applications this is not feasible. Therefore, to consider only backward-looking features is a significant improvement toward making the method useful in practical applications. We leave it for future research to consider feature selection in greater depth.

Model Fitting

Exhibit 2 outlines the jump model approach to estimating an HMM given a batch of time-series data (Nystrup, Lindström, and Madsen 2020b). After constructing a set of standardized features, the algorithm iterates between (a) finding the parameters that minimize the loss function for a given state sequence (i.e., fitting the model) and (b) finding the state sequence that minimizes the objective function (Equation 1) given the parameters (i.e., fitting the state sequence). This process is repeated until the state sequence does not change. In practice, we finish either after a maximum of 10 iterations or once the value of the objective function changes by less than 10^{-6} from one iteration to the next. Convergence usually is reached within fewer than five iterations. Note that although we focus on the estimation of HMMs, the jump model framework is more general and can be used to fit a much wider range of models.

Exhibit 1
Features for HMM Estimation

Input: Time series $y = \{y_1, \ldots, y_t\}$ and window length l

1. Observation: y_t
2. Absolute change: $|y_t - y_{t-1}|$
3. Previous absolute change: $|y_{t-1} - y_{t-2}|$
4. Centered mean: mean $[y_{t-l+1}, \ldots, y_t]$
5. Centered standard deviation: std $[y_{t-l+1}, \ldots, y_t]$
6. Left mean: mean $\left[y_{t-l+1}, \ldots, y_{t-\frac{l}{2}}\right]$
7. Left standard deviation: std $\left[y_{t-l+1}, \ldots, y_{t-\frac{l}{2}}\right]$
8. Right mean: mean $\left[y_{t-\frac{l}{2}+1}, \ldots, y_t\right]$
9. Right standard deviation: std $\left[y_{t-\frac{l}{2}+1}, \ldots, y_t\right]$

Output: Feature set

To improve the quality of the solution, we run the algorithm in Exhibit 2 from 10 different initial state sequences generated randomly using K-means++ (Arthur and Vassilvitskii 2007) and keep the model that achieves the lowest objective value. Based on the final estimate of the state sequence s^i, transition probabilities are computed by counting the number of transitions between states along with the distributional parameters for each state.

In Exhibit 2, the parameters $\theta^i = \{\theta_1^i, \ldots, \theta_K^i\}$ are estimated by computing analytically the minimum of the loss function for each of the K states. Subsequently, the most likely sequence of states is found using dynamic programming (Bellman 1957). Define

$$V(T, s) = \ell(z_T, \theta_s), \tag{2}$$

$$V(t, i) = \ell(z_t, \theta_i) + \min_j \{V(t+1, j) + \lambda \mathbb{I}_{i \neq j}\}, \quad t = T-1, \ldots, 1. \tag{3}$$

The most likely sequence of states then is given by

$$s_1 = \underset{j}{\operatorname{argmin}}\, V(1, j), \tag{4}$$

$$s_t = \underset{j}{\operatorname{argmin}} \{V(t, j) + \lambda \mathbb{I}_{s_{t-1} \neq j}\}, \quad t = 2, \ldots, T. \tag{5}$$

EXHIBIT 2
Jump Estimation of HMM

Input: Time series $y = \{y_1, \ldots, y_T\}$, number of latent states K, jump penalty λ, and initial state sequence $s^0 = \{s_1^0, \ldots, s_T^0\}$

1. Construct a set of standardized features z from the time series y
2. Iterate for $i = 1, \ldots$
   (a) Fit model $\theta^i = \text{argmin}_\theta \sum_{t=1}^{T} \ell\left(z_t, \theta_{s_t^{i-1}}\right)$
   (b) Fit state sequence $s^i = \text{argmin}_s \left\{ \sum_{t=1}^{T-1} \left[\ell\left(z_t, \theta_{s_t}^i\right) + \lambda \, \mathbb{I}_{s_t \neq s_{t+1}} \right] + \ell\left(z_T, \theta_{s_T}^i\right) \right\}$
3. Until $s^i = s^{i-1}$
4. Compute the transition probabilities and distributional parameters for each state

Output: HMM parameters and prediction of latent states

Note that if the time order of operations is reversed, Equations 2 to 5 become the Viterbi algorithm (Viterbi 1967).

Complexity. Finding the most likely sequence of states requires $O(TK^2)$ operations, which is the same as one forward and backward pass by the EM algorithm. Although the jump estimator and EM algorithms have the same computational complexity per iteration, the jump estimator converges faster. In our simulation study, typically the jump estimator requires about five iterations, whereas the EM algorithm needs between 50 and 100 iterations to converge.

Greedy Online State Classification

In many practical applications, it is of critical significance to estimate the model recursively in an online fashion. An advantage of our jump estimator is that it is easily adaptable to online settings. Exhibit 3 outlines how states can be inferred recursively once an initial jump model has been fit using the algorithm in Exhibit 2. It is a greedy algorithm because it estimates which state an observation belongs to without parsing historical observations. Without the fixed-cost jump penalty, a greedy approach would be expected to yield less persistent state estimates.

The model parameters θ in Exhibit 3 are the same as in Exhibit 2 (i.e., the state-dependent means of the feature vectors and not the HMM parameters). Importantly, our greedy online state classifier is not based on any assumptions about the conditional or sojourn-time distributions. Although not explored in this article, we note that each time the arrival cost has been updated and the state computed, the model parameters can be updated in a straightforward way. In addition, the parameter update can easily be modified to incorporate fading memory to put greater emphasis on more recent observations (Nystrup et al. 2019). If the greedy algorithm is combined with a parameter update each time a state is estimated and the arrival cost updated, it becomes an online clustering approach.

EXHIBIT 3
Greedy Online State Classification

Input: Model parameters θ, jump penalty λ, last two observations $\{z_{t-1}, z_t\}$, and arrival cost $\mathcal{A}_{t-1}$

1. Update $\mathcal{A}_t(s_t) = \min_{s_{t-1}} \left\{ \ell(z_{t-1}, \theta_{s_{t-1}}) + \mathcal{A}_{t-1}(s_{t-1}) + \lambda \mathbb{I}_{s_{t-1} \neq s_t} \right\}$
2. Compute $\hat{s}_t = \text{argmin}_s \{ \ell(z_t, \theta_s) + \mathcal{A}_t(s) \}$

Output: Estimated state $\hat{s}_t$ and updated arrival cost $\mathcal{A}_t$

Another advantage of our classifier is its lower memory requirements. Compared with the Viterbi algorithm, which requires storing and parsing of all historical observations, by estimating the states greedily without parsing historical (or later) observations, our classifier only needs to store the last feature vector and arrival cost. This is a great benefit, especially in real-time trading applications and in memory impoverished systems, such as Internet of Things devices (Kumar, Goyal, and Varma 2017; Zhu et al. 2019).

SIMULATION STUDY

We compare the accuracy of HMMs estimated using maximum likelihood, spectral clustering, and jump estimation in simulations. The advantage of a

simulation study is that the true state sequence is known, which makes it possible to evaluate the ability of each model to correctly identify the underlying hidden states.

We define accuracy as the maximum classification accuracy among all possible alignments of an inferred state sequence. Specifically, accuracy is the ratio $tp/(tp + fn)$, where tp is the number of true positives and fn the number of false negatives. It is necessary to check whether a different permutation of the states yields higher accuracy owing to the risk of label switching.

In our simulations, we report the balanced accuracy (BAC), which is the average accuracy per observed state, to avoid inflated performance estimates on imbalanced datasets (Brodersen et al. 2010).

Two-State HMM

Following Zheng, Li, and Xu (2019) and Nystrup, Lindström, and Madsen (2020b), we simulate data from the two-state Gaussian HMM

$$y_t | s_t \sim N(\mu_{s_t}, \sigma^2_{s_t}), \tag{6}$$

where s_t is a first-order Markov chain, with parameters

$$\begin{aligned} &\mu_1 = .0006, \qquad \mu_2 = -.0008, \\ &\sigma_1 = .0078, \qquad \sigma_2 = .0174, \\ &\Gamma = \begin{pmatrix} .9979 & .0021 \\ .0120 & .9880 \end{pmatrix}. \end{aligned} \tag{7}$$

These parameters are the daily equivalent of a model Hardy (2001) estimated from monthly returns on a stock market index. State one is the most persistent and has the lowest variance and a positive mean. State two has a much higher variance and a negative mean value. This model captures the stylized behavior of many financial time series, including volatility clustering and leptokurtosis (Rydén, Teräsvirta, and Åsbrink 1998). The significant overlap among the conditional distributions makes it challenging to infer the unobserved states.

Selecting λ

Exhibit 4 shows the BAC of the jump estimator as a function of the penalty parameter λ for different sample lengths when applying the algorithm in Exhibit 2 to batches of data simulated from Equation 7. Specifically, it depicts the trade-off between accuracy and persistence. The main difference between sample lengths is the level of BAC, with longer series leading to higher BAC. When the value of the penalty parameter is too large, the BAC approaches 1/2. When the penalty parameter is too small, the BAC converges to that of the unpenalized case, which corresponds to applying K-means clustering to the feature matrix. We find that across different sample lengths, the optimal value of the penalty parameter is around $\lambda = 100$. By comparison, the optimal value is around $\lambda = 50$ when applying the greedy online classifier in Exhibit 3 based on the estimated jump model parameters.

In-Sample Results

Exhibit 5 shows the mean (standard deviation) of the estimated parameters when applying the different estimators in sample to 1,000 simulated series of different lengths. The accuracy using the true parameters is based on the Viterbi path. The reported maximum likelihood estimates (MLEs) are obtained using the regularized EM algorithm (Huang, Acero, and Hon 2001, Chapter 9.6.1) that is implemented in the hmmlearn package in Python with the variance prior set to 10^{-4}. The MLE is initialized using K-means with 10 different centroid seeds generated using K-means++ (Arthur and Vassilvitskii 2007), similar to how spectral clustering and jump estimation are initialized. The accuracy of the MLEs could be improved by providing initial values that are close to the true values, but that would not be a fair comparison. Compared with the results reported by Nystrup, Lindström, and Madsen (2020b) using a mix of backward- and forward-looking features, the BACs in Exhibit 5 are only a few percentage points lower.

With a sample size of only 250 observations, the jump estimation approach outlined in Exhibit 2 with $\lambda = 100$ is better than MLE and spectral clustering at estimating nearly all parameters and states. Notably, the biggest difference is in the estimates of the transition probabilities, with MLE and spectral clustering significantly overestimating them, whereas the jump estimates are close to the true values.

With a sample size of 500 observations, the BAC of the jump estimate is still higher than that of the MLE, with spectral clustering far behind. Although the MLEs of the transition probabilities have improved, they are still far from the true values. Even with a sample size of

EXHIBIT 4

Balanced Accuracy of the Jump Estimator as a Function of the Penalty Parameter for Different Sample Lengths

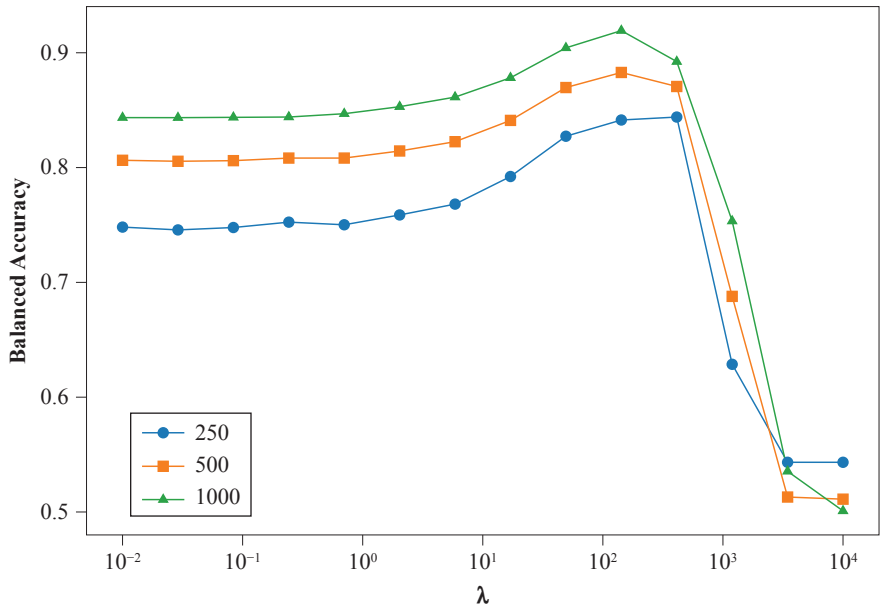


EXHIBIT 5

Parameter Estimates Based on 1,000 Simulated Series of Different Lengths

	μ_1	μ_2	σ_1	σ_2	γ_{12}	γ_{21}	ACCURACY 1	ACCURACY 2	BAC
True	0.0006	−0.0008	0.0078	0.0174	0.0021	0.0120	0.9965	0.8752	0.9536
							(0.0146)	(0.2339)	(0.1027)
250									
MLE	0.0006	−0.0001	0.0074	0.0117	0.2884	0.3707	0.7747	**0.8290**	0.7517
	(0.0021)	(0.0039)	(0.0014)	(0.0048)	(0.2418)	(0.2853)	(0.2029)	(0.2351)	(0.1879)
Spec	0.0006	−0.0001	0.0075	0.0118	0.0997	0.1612	0.7175	0.7589	0.6924
	(0.0018)	(0.0035)	(0.0018)	(0.0046)	(0.0535)	(0.0589)	(0.1874)	(0.1890)	(0.1526)
Jump	0.0006	−0.0001	**0.0079**	**0.0120**	**0.0055**	**0.0242**	**0.8611**	0.8263	**0.8303**
	(0.0009)	(0.0031)	(0.0014)	(0.0049)	(0.0062)	(0.0277)	(0.1460)	(0.2398)	(0.1457)
500									
MLE	0.0007	−0.0003	0.0075	0.0139	0.1805	0.2289	0.8648	**0.8647**	0.8293
	(0.0012)	(0.0023)	(0.0009)	(0.0046)	(0.2314)	(0.2612)	(0.1892)	(0.2074)	(0.1852)
Spec	0.0007	**−0.0005**	0.0075	0.0137	0.0773	0.1463	0.7913	0.7728	0.7562
	(0.0014)	(0.0031)	(0.0012)	(0.0045)	(0.0559)	(0.0604)	(0.1867)	(0.1895)	(0.1549)
Jump	**0.0006**	−0.0003	**0.0079**	**0.0143**	**0.0034**	**0.0204**	**0.9198**	0.8458	**0.8736**
	(0.0005)	(0.0028)	(0.0005)	(0.0046)	(0.0024)	(0.0206)	(0.1239)	(0.2095)	(0.1382)
1,000									
MLE	**0.0006**	−0.0005	0.0076	0.0154	0.0962	0.1382	0.9365	**0.8846**	0.8961
	(0.0006)	(0.0017)	(0.0004)	(0.0038)	(0.1799)	(0.2140)	(0.1327)	(0.1776)	(0.1447)
Spec	0.0008	−0.0010	0.0076	0.0149	0.0534	0.1325	0.8521	0.7936	0.8140
	(0.0009)	(0.0024)	(0.0008)	(0.0039)	(0.0463)	(0.0568)	(0.1599)	(0.1630)	(0.1248)
Jump	**0.0006**	**−0.0007**	**0.0078**	**0.0160**	**0.0025**	**0.0188**	**0.9669**	0.8727	**0.9173**
	(0.0003)	(0.0023)	(0.0003)	(0.0034)	(0.0015)	(0.0163)	(0.0727)	(0.1509)	(0.0941)

Notes: The reported values are the mean (standard deviation) of the estimated parameters. Bold entries identify the best estimate among MLE, spectral clustering, and jump estimation.

EXHIBIT 6

Parameter Estimates Based on 1,000 Series of Length 500 Simulated from Conditional t_5 Distributions

	μ_1	μ_2	σ_1	σ_2	γ_{12}	γ_{21}	ACCURACY 1	ACCURACY 2	BAC
True	0.0006	−0.0008	0.0078	0.0174	0.0021	0.0120	0.9869	0.8238	0.9291
							(0.0228)	(0.2752)	(0.1231)
MLE	**0.0006**	−0.0011	0.0066	0.0166	0.1175	0.4149	0.9260	0.6936	0.8368
	(0.0005)	(0.0160)	(0.0021)	(0.0047)	(0.1341)	(0.3488)	(0.1102)	(0.2931)	(0.1456)
Spec	0.0007	−0.0005	**0.0076**	0.0147	0.0565	0.1467	0.8023	0.6887	0.7481
	(0.0016)	(0.0040)	(0.0031)	(0.0059)	(0.0398)	(0.0646)	(0.1634)	(0.2424)	(0.1351)
Jump	**0.0006**	**−0.0006**	**0.0080**	**0.0171**	**0.0043**	**0.0386**	**0.9374**	**0.7432**	**0.8587**
	(0.0016)	(0.0054)	(0.0037)	(0.0088)	(0.0106)	(0.0427)	(0.1216)	(0.2941)	(0.1497)

Notes: The reported values are the mean (standard deviation) of the estimated parameters. Bold entries identify the best estimate among MLE, spectral clustering, and jump estimation.

1,000 observations, neither spectral clustering nor MLE is able to capture the true persistence of the underlying state process. The BAC of MLE is slightly lower than that of the jump estimator, and both methods are fairly close to the accuracy of the Viterbi path based on the true parameter values.

Of course, if we kept increasing the sample size, the performance of MLE would keep improving. However, it is not realistic for many practical purposes to have more than 1,000 daily observations that are representative of current conditions available for estimation.

Misspecified Conditional Distributions

To compare the different estimation methods when conditional distributions are misspecified, we repeat the simulation study, but this time we sample the observations from Student's *t* rather than Gaussian distributions. The conditional distributions have the same mean and standard deviation but are *t*-distributions with five degrees of freedom, which have substantially heavier tails than the Gaussian distribution. Several studies have found that this is a better fit to the empirical distribution of financial returns (Cont 2001; Bulla 2011; Nystrup, Madsen, and Lindström 2017).

Exhibit 6 shows the result when simulating 1,000 series of 500 observations from Student's *t*-distributions. The BACs are about the same as in Exhibit 5. Sampling from conditional distributions with heavier tails significantly inflates the transition probability estimates by MLE, leading to an estimate of 41% compared with the true value of 1% probability of jumping from state two to one. Both the spectral clustering and jump estimators are far more robust to the misspecified conditional distributions.

Misspecified Sojourn-Time Distributions

To compare the different estimation methods when sojourn-time distributions are misspecified, we repeat the simulation study, but now we sample the sojourn times from negative binomial distributions instead of geometric distributions. The negative binomial distributions are calibrated to yield the same expected sojourn duration as before by adjusting the probability parameters. We fix the shape parameter to $n_1 = 0.1$ in state one and $n_2 = 0.06$ in state two based on the values Bulla and Bulla (2006) reported from fitting similar hidden semi-Markov models to daily stock returns. This leads to an increased number of very short sojourns and prolonged sojourns.

Exhibit 7 shows the result when sampling from misspecified sojourn-time distributions. The negative binomial distribution cannot be summarized by a single transition probability parameter. However, the inverse of the true and estimated transition probabilities corresponds to the average sojourn duration. More so than when sampling from misspecified conditional distributions, the accuracy of the MLEs deteriorates significantly compared with Exhibit 5, whereas the jump estimator is remarkably robust. The jump estimator maintains a high accuracy for both states and for the sojourn duration.

EXHIBIT 7

Parameter Estimates Based on 1,000 Series of Length 500 Simulated from Negative Binomial Sojourn-Time Distributions

	μ_1	μ_2	σ_1	σ_2	γ_{12}	γ_{21}	ACCURACY 1	ACCURACY 2	BAC
True	0.0006	−0.0008	0.0078	0.0174	0.0021	0.0120	0.9739	0.8329	0.9375
							(0.1369)	(0.3272)	(0.1449)
MLE	0.0003	0.0001	**0.0084**	0.0123	0.3085	0.3449	0.7682	0.7880	0.7396
	(0.0027)	(0.0032)	(0.0031)	(0.0045)	(0.2417)	(0.2534)	(0.2165)	(0.2551)	(0.1998)
Spec	**0.0006**	**−0.0005**	0.0087	0.0126	0.1072	0.1517	0.7191	0.7437	0.6943
	(0.0023)	(0.0030)	(0.0038)	(0.0044)	(0.0579)	(0.0558)	(0.2264)	(0.2185)	(0.1756)
Jump	0.0005	−0.0002	0.0088	**0.0129**	**0.0053**	**0.0189**	**0.8663**	**0.8434**	**0.8420**
	(0.0018)	(0.0025)	(0.0031)	(0.0047)	(0.0088)	(0.0232)	(0.1858)	(0.2510)	(0.1648)

Notes: The reported values are the mean (and standard deviation) of the estimated parameters. Bold entries identify the best estimate among MLE, spectral clustering, and jump estimation.

Out-of-Sample Results

Next, we consider an out-of-sample application in which observations are processed sequentially one by one. We start with the Gaussian models that were fit earlier based on 1,000 series of different lengths simulated using Equation 7. We apply these models sequentially to subsequent samples of 250 observations without re-estimating any parameters. Because spectral clustering was not able to contend with MLE and the jump estimator in sample, we choose to focus on the latter two in the out-of-sample study.

Online state classification without latency is a significantly harder task than in-sample state inference on a batch of data to which filtering algorithms can be applied to smooth the estimated state sequence. Based on the in-sample MLE, we apply the Viterbi algorithm sequentially to the out-of-sample test data, every time outputting the state corresponding to the last observation. For the jump models, we apply the algorithm in Exhibit 3 with $\lambda = 50$ to greedily infer the states underlying the out-of-sample test data. It is natural to expect that a greedy approach will lead to lower classification accuracy compared with an approach that parses historical observations before classifying each new observation, but that is not the case in our study.

Exhibit 8 shows the result when inferring the hidden states sequentially on 1,000 simulated series of length 250 based on the models estimated in sample from samples of different length. The reported transition probabilities are based on counting the number of transitions, whereas the means and variances are estimated conditional on the classified state sequences. Compared with Exhibit 5, the BACs are two to four percentage points lower. Interestingly, both estimators have a much harder time identifying state two out of sample compared with in sample.

Similar to the in-sample study, the most significant difference between the MLE-based and greedy state classifications is the estimated transition frequencies. In the out-of-sample test, the implied transition probabilities based on the MLE state sequences are significantly worse. Although the greedy estimates of the transition probabilities are higher than the in-sample jump estimates, they are still much closer to the true values.

Intraday Data

The results presented in Exhibit 8 made us wonder what could be done to improve the greedy classifier's ability to identify observations belonging to the high-volatility state. Aside from its robustness and ability to estimate persistent state processes, the biggest advantage of the jump estimator compared with MLE is its potential for easily extending the feature space.

As a simple illustration, we explore how accuracy can be improved when observations sampled at a frequency higher than daily are available for estimation of the standard deviation features in Exhibit 1. In financial applications, it is common to use intraday data to improve estimates of daily volatility (Barndorff-Nielsen and Shephard 2002).

EXHIBIT 8

Parameters Estimated Online Out of Sample Based on 1,000 Simulated Series When Different Sample Lengths Were Available for In-Sample Estimation

	μ_1	μ_2	σ_1	σ_2	γ_{12}	γ_{21}	ACCURACY 1	ACCURACY 2	BAC
True	0.0006	−0.0008	0.0078	0.0174	0.0021	0.0120	0.9695	0.8014	0.9367
							(0.0743)	(0.2767)	(0.1146)
250									
MLE	0.0005	0.0003	0.0087	0.0107	0.2455	0.4910	0.7007	**0.7099**	0.7175
	(0.0010)	(0.0018)	(0.0019)	(0.0038)	(0.2353)	(0.3031)	(0.2327)	(0.2408)	(0.1858)
Jump	0.0005	**0.0002**	**0.0084**	**0.0114**	**0.0153**	**0.0676**	**0.8420**	0.6635	**0.8020**
	(0.0011)	(0.0020)	(0.0014)	(0.0043)	(0.0358)	(0.1001)	(0.1797)	(0.3457)	(0.1777)
500									
MLE	0.0005	0.0001	0.0085	0.0119	0.1491	0.3852	0.7923	**0.7779**	0.8060
	(0.0011)	(0.0019)	(0.0017)	(0.0042)	(0.2123)	(0.3226)	(0.2300)	(0.2477)	(0.1936)
Jump	**0.0006**	**−0.0001**	**0.0084**	**0.0128**	**0.0109**	**0.0606**	**0.8976**	0.7120	**0.8586**
	(0.0008)	(0.0025)	(0.0016)	(0.0044)	(0.0178)	(0.0906)	(0.1627)	(0.3209)	(0.1684)
1,000									
MLE	0.0006	0.0000	0.0083	0.0131	0.0687	0.2743	0.8563	**0.8408**	0.8678
	(0.0007)	(0.0024)	(0.0012)	(0.0044)	(0.1391)	(0.3123)	(0.2128)	(0.2307)	(0.1633)
Jump	0.0006	**−0.0001**	**0.0082**	**0.0141**	**0.0072**	**0.0611**	**0.9487**	0.7216	**0.8953**
	(0.0006)	(0.0026)	(0.0011)	(0.0044)	(0.0134)	(0.1130)	(0.0992)	(0.2995)	(0.1432)

Notes: The out-of-sample testing is done on samples of length 250. The reported values are the mean (standard deviation) of the estimated parameters. Bold entries identify the best estimate between the MLE-based Viterbi states and the greedy jump states.

Exhibit 9 shows the result when observations sampled at a higher frequency are available for estimation of the standard deviation features. It is apparent that sampling at frequencies of 2, 5, and 10 observations per day significantly improves the greedy classifier's accuracy for state two. Compared with Exhibit 8, when two observations per day are available for feature estimation, the improvement in BAC is about one percentage point. With 5 and 10 observations per day available for feature estimation, the improvements in BAC are two and three percentage points, respectively. By sampling 10 times per day, the BAC of the greedy classifier is one percentage point higher than the in-sample BAC in Exhibit 5. This result is very encouraging considering how much harder online out-of-sample state classification is. In addition, we note that increasing the sampling frequency improves estimates of the transition probabilities compared with Exhibit 8.

APPLICATION TO S&P 500

Finally, we consider an application to the S&P 500 Index. We consider daily log returns along with five-minute subsampled realized volatility data covering the period from January 2000 through March 2020.[1] First, we use the data from January 2000 to January 2005 to fit a jump model with $\lambda = 1{,}000$ using the algorithm in Exhibit 2. Second, we use the data from January 2005 to January 2010 to select $\lambda = 500$ to use in the greedy online classifier in Exhibit 3. Third, we use the greedy online classifier to estimate states for the period from January 2010 through March 2020.

Exhibit 10 shows the S&P 500 Index and its annualized realized volatility in percentage (five-minute subsampled) on the test period from January 2010 through March 2020, along with three different state sequences. The first two state sequences are estimated using the greedy online classifier in Exhibit 3 based on jump models with and without the realized volatility features. The third state sequence is estimated by applying the Viterbi algorithm sequentially to the test data, based on an HMM fitted to the log returns from January 2000 to January 2010 using maximum likelihood estimation.

[1] The five-minute subsampled realized volatility data for the S&P 500 Index were downloaded from https://realized.oxford-man.ox.ac.uk/data/download.

EXHIBIT 9

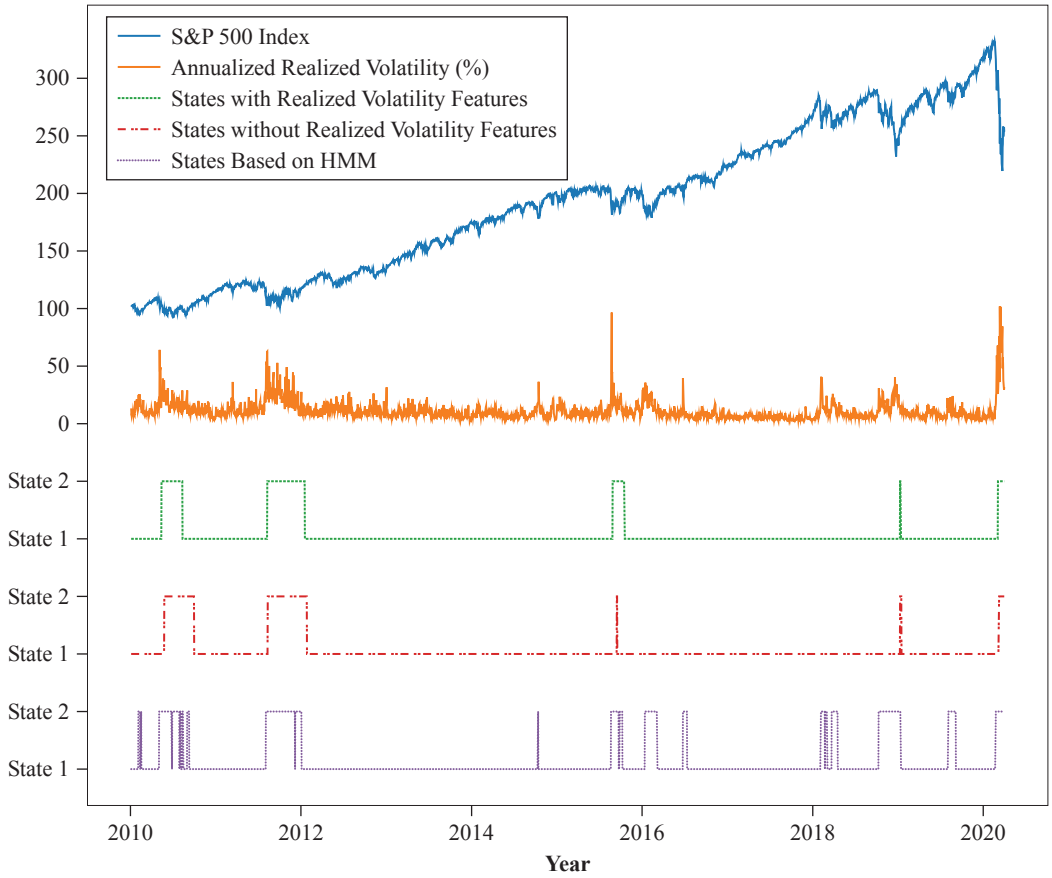
Parameters Estimated Using the Greedy Online Classifier Out of Sample Based on 1,000 Simulated Series When Observations Sampled at Higher Frequencies Were Available for Estimation of the Volatility Features

	μ_1	μ_2	σ_1	σ_2	γ_{12}	γ_{21}	ACCURACY 1	ACCURACY 2	BAC
True	0.0006	−0.0008	0.0078	0.0174	0.0021	0.0120	0.9695	0.8014	0.9367
							(0.0743)	(0.2767)	(0.1146)
2	0.0006	0.0000	0.0082	0.0128	0.0106	0.0555	0.9020	0.7508	0.8676
	(0.0007)	(0.0025)	(0.0012)	(0.0045)	(0.0361)	(0.1004)	(0.1567)	(0.3327)	(0.1693)
5	0.0006	−0.0001	0.0081	0.0130	0.0088	0.0459	0.9029	0.7760	0.8770
	(0.0007)	(0.0025)	(0.0009)	(0.0044)	(0.0195)	(0.0864)	(0.1595)	(0.3315)	(0.1704)
10	0.0006	0.0000	0.0080	0.0131	0.0082	0.0528	0.9068	0.7991	0.8856
	(0.0007)	(0.0022)	(0.0010)	(0.0044)	(0.0173)	(0.1148)	(0.1587)	(0.3131)	(0.1593)

Notes: Each series consists of 500 observations for in-sample estimation and 250 observations for out-of-sample testing. The reported values are the mean (standard deviation) of the estimated parameters.

EXHIBIT 10

Comparison of State Sequences Estimated Based on the S&P 500 Index and Its Realized Volatility



Note: The legends are sorted according to the order of the time series.

Adding two extra features, which are moving averages with window lengths $l = 6$ and $l = 14$ of the five-minute subsampled realized volatility observations, leads to improved state estimates. In particular, the greedy classifier can detect state changes slightly faster based on the realized volatility features. Although the differences are measured in days and not months, they can still make a substantial difference because they happen during the most turbulent times.

Comparing the two state sequences that are based on jump models with that based on an HMM, the differences in persistence are striking. The HMM detects changes as quickly as the jump model with the realized volatility features; however, this comes at the cost of many spurious and short-lived state changes. If the estimated states were used as the basis of a regime-based asset allocation strategy, the lower persistence would lead to substantially higher turnover, which could easily make a difference in whether the strategy is profitable or not (Nystrup et al. 2015; Nystrup et al. 2017; Nystrup, Madsen, and Lindström 2018). The results from this application suggest that using data sampled at higher frequencies can help improve the classification of persistent states in financial time series. Determination of the optimal sampling frequency is outside the scope of this article (Ait-Sahalia, Mykland, and Zhang 2005).

CONCLUSIONS

We propose a greedy online classifier that contemporaneously determines which hidden state a new observation belongs to without the need to parse historical observations and without compromising persistence. We believe this classifier will be useful in various practical applications in which it is important to classify time-series data without any latency while maintaining persistence in the identified states. Through a series of simulations, we showed that (1) in most settings the new classifier remarkably obtains a higher accuracy than the correctly specified maximum likelihood estimator and (2) the new classifier is more robust to misspecification, yielding state sequences that are significantly more persistent both in and out of sample. Additionally, we demonstrated how classification accuracy can be improved by including features that are based on data sampled at higher frequencies.

The MLE required at least 1,000 observations to compete with the proposed jump estimator based on backward-looking temporal features and penalizing jumps, a strong argument in favor of the jump estimator. Furthermore, we demonstrated that the jump estimates of the transition probabilities were significantly more precise. In the sequential out-of-sample test, in which it was harder to obtain persistent state classifications because of the inability to apply smoothing, the differences between the states obtained from MLE and the jump model were even more pronounced.

The robustness of the jump estimator to initialization, misspecification, and data scarcity should make it the preferred choice for practical applications in which, by definition, all models are wrong. With the same computational complexity per iteration as the EM algorithm, it is a significant advantage of the jump estimator that it needs 10 to 20 times fewer iterations to converge. The computational advantage of the jump estimator will be particularly useful in high-frequency applications. As demonstrated in the simulation study, higher sampling rates can lead to accuracy gains.

Future Work

In closing, we comment on some directions for future work. In many situations, the researcher has access to application-specific exogenous data that may improve state classification if added to the feature space. For example, for time series of security returns and changes in macroeconomic variables, we hypothesize that news- and sentiment-based features could be useful. In these situations, it is natural to consider approaches for determining the appropriate features to include. In a future article, we will explore feature selection techniques in the jump estimation framework.

We anticipate that a generalization of the framework to multivariate HMMs is straightforward, using the features in Exhibit 1 as well as features that exploit the dependence between variables. We believe that the robustness of the jump framework will be beneficial in clustering high-dimensional time series with large feature sets.

In future work, we plan to leverage the increased persistence of the jump model in regime-based asset allocation strategies.

ACKNOWLEDGMENTS

This work was supported by Vergstiftelsen and the Centre for IT-Intelligent Energy Systems (CITIES) project, funded in part by Innovation Fund Denmark under grant no. 1305-00027B.

REFERENCES

Ait-Sahalia, Y., P. A. Mykland, and L. Zhang. 2005. "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise." *The Review of Financial Studies* 18 (2): 351–416.

Andersson, S., T. Rydén, and R. Johansson. 2003. "Linear Optimal Prediction and Innovations Representations of Hidden Markov Models." *Stochastic Processes and Their Applications* 108 (1): 131–149.

Arthur, D., and S. Vassilvitskii. 2007. "k-Means++: The Advantages of Careful Seeding." In *Proceedings of the Eighteenth Annual ACM–SIAM Symposium on Discrete Algorithms*, pp. 1027–1035. Philadelphia: Society for Industrial and Applied Mathematics, 2007.

Barndorff-Nielsen, O. E., and N. Shephard. 2002. "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (2): 253–280.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss. 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *Annals of Mathematical Statistics* 41 (1): 164–171.

Beal, M. J., Z. Ghahramani, and C. E. Rasmussen. 2002. "The Infinite Hidden Markov Model." *Advances in Neural Information Processing Systems* 14: 577–584.

Bellman, R. E. *Dynamic Programming*. Princeton, NJ: Princeton University Press. 1957.

Bemporad, A., V. Breschi, D. Piga, and S. Boyd. 2018. "Fitting Jump Models." *Automatica* 96: 11–21.

Bhardwaj, S., V. Sharma, S. Srivastava, O. Sastry, B. Bandyopadhyay, S. Chandel, and J. Gupta. 2013. "Estimation of Solar Radiation Using a Combination of Hidden Markov Model and Generalized Fuzzy Model." *Solar Energy* 93: 43–54.

Brodersen, K. H., C. S. Ong, K. E. Stephan, and J. M. Buhmann. "The Balanced Accuracy and Its Posterior Distribution." In *20th International Conference on Pattern Recognition*, pp. 3121–3124. Istanbul: IEEE. 2010.

Bulla, J. 2011. "Hidden Markov Models with t Components. Increased Persistence and Other Aspects." *Quantitative Finance* 11 (3): 459–475.

Bulla, J., and I. Bulla. 2006. "Stylized Facts of Financial Time Series and Hidden Semi-Markov Models." *Computational Statistics & Data Analysis* 51 (4): 2192–2209.

Cont, R. 2001. "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues." *Quantitative Finance* 1 (2): 223–236.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1): 1–38.

Dias, J. G., and S. B. Ramos. 2014. "Dynamic Clustering of Energy Markets: An Extended Hidden Markov Approach." *Expert Systems with Applications* 41 (17): 7722–7729.

Fiecas, M., J. Franke, R. von Sachs, and J. T. Kamgaing. 2017. "Shrinkage Estimation for Multivariate Hidden Markov Models." *Journal of the American Statistical Association* 112 (517): 424–435.

Fox, E. B., E. B. Sudderth, M. I. Jordan, and A. S. Willsky. 2011. "A Sticky HDP-HMM with Application to Speaker Diarization." *Annals of Applied Statistics* 5 (2A): 1020–1056.

Gales, M., and S. Young. 2008. "The Application of Hidden Markov Models in Speech Recognition." *Foundations and Trends in Signal Processing* 1 (3): 195–304.

Garg, V., and T. Pichkhadze. 2019. "Online Markov Decoding: Lower Bounds and Near-Optimal Approximation Algorithms." In *Advances in Neural Information Processing Systems* 32: 5681–5691.

Ghahramani, Z., and M. I. Jordan. 1997. "Factorial Hidden Markov Models." *Machine Learning* 29 (2–3): 245–273.

Hallac, D., P. Nystrup, and S. Boyd. 2019. "Greedy Gaussian Segmentation of Multivariate Time Series." *Advances in Data Analysis and Classification* 13 (3): 727–751.

Hardy, M. R. 2001. "A Regime-Switching Model of Long-Term Stock Returns." *North American Actuarial Journal* 5 (2): 41–53.

Huang, X., A. Acero, and H. W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River, NJ: Prentice-Hall. 2001.

Kang, M., J. Ahn, and K. Lee. 2018. "Opinion Mining Using Ensemble Text Hidden Markov Models for Text Classification." *Expert Systems with Applications* 94: 218–227.

Kim, S. J., K. Koh, S. Boyd, and D. Gorinevsky. 2009. "l_1 Trend Filtering." *SIAM Review* 51 (2): 339–360.

Kumar, A., S. Goyal, and M. Varma. 2017. "Resource-Efficient Machine Learning in 2 KB RAM for the Internet of Things." *Proceedings of the 34th International Conference on Machine Learning* 70: 1935–1944.

Langrock, R., and W. Zucchini. 2011. "Hidden Markov Models with Arbitrary State Dwell-Time Distributions." *Computational Statistics & Data Analysis* 55 (1): 715–724.

Lloyd, S. P. 1982. "Least Squares Quantization in PCM." *IEEE Transactions on Information Theory* 28 (2): 129–137.

Narasimhan, M., P. Viola, and M. Shilman. 2006. "Online Decoding of Markov Models under Latency Constraints." *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, pp. 657–664.

Nefian, A. V., and M. H. Hayes. 1988. "Hidden Markov Models for Face Recognition." *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing* 5: 2721–2724.

Ng, A. Y., M. I. Jordan, and Y. Weiss. 2002. "On Spectral Clustering: Analysis and an Algorithm." *Advances in Neural Information Processing Systems* 14: 849–856.

Nystrup, P., S. Boyd, E. Lindström, and H. Madsen. 2019. "Multi-Period Portfolio Selection with Drawdown Control." *Annals of Operations Research* 282 (1–2): 245–271.

Nystrup, P., B. W. Hansen, H. O. Larsen, H. Madsen, and E. Lindström. 2017. "Dynamic Allocation or Diversification: A Regime-Based Approach to Multiple Assets." *The Journal of Portfolio Management* 44 (2): 62–73.

Nystrup, P., B. W. Hansen, H. Madsen, and E. Lindström. 2015. "Regime-Based versus Static Asset Allocation: Letting the Data Speak." *The Journal of Portfolio Management* 42 (1): 103–109.

——. 2016. "Detecting Change Points in VIX and S&P 500: A New Approach to Dynamic Asset Allocation." *Journal of Asset Management* 17 (5): 361–374.

Nystrup, P., E. Lindström, and H. Madsen. 2020a. "Hyperparameter Optimization for Portfolio Selection." *The Journal of Financial Data Science* 2 (3).

——. 2020b. "Learning Hidden Markov Models with Persistent States by Penalizing Jumps." *Expert Systems with Applications* 150: 113307.

Nystrup, P., H. Madsen, and E. Lindström. 2015. "Stylised Facts of Financial Time Series and Hidden Markov Models in Continuous Time." *Quantitative Finance* 15 (9): 1531–1541.

——. 2017. "Long Memory of Financial Time Series and Hidden Markov Models with Time Varying Parameters." *Journal of Forecasting* 36 (8): 989–1002.

——. 2018. "Dynamic Portfolio Optimization across Hidden Market Regimes." *Quantitative Finance* 18 (1): 83–95.

Petropoulos, A., S. P. Chatzis, and S. Xanthopoulos. 2016. "A Novel Corporate Credit Rating System Based on Student's-*t* Hidden Markov Models." *Expert Systems with Applications* 53: 87–105.

Pinson, P., L. Christensen, H. Madsen, P. Sørensen, M. Donovan, and L. Jensen. 2008. "Regime-Switching Modelling of the Fluctuations of Offshore Wind Generation." *Journal of Wind Engineering and Industrial Aerodynamics* 96 (12): 2327–2347.

Robinson, W. N., and A. Aria. 2018. "Sequential Fraud Detection for Prepaid Cards Using Hidden Markov Model Divergence." *Expert Systems with Applications* 91: 235–251.

Rousseeuw, K., É. P. Caillault, A. Lefebvre, and D. Hamad. 2015. "Hybrid Hidden Markov Model for Marine Environment Monitoring." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8 (1): 204–213.

Rydén, T. 2008. "EM versus Markov Chain Monte Carlo for Estimation of Hidden Markov Models: A Computational Perspective." *Bayesian Analysis* 3 (4): 659–688.

Rydén, T., T. Teräsvirta, and S. Åsbrink. 1998. "Stylized Facts of Daily Return Series and the Hidden Markov Model." *Journal of Applied Econometrics* 13 (3): 217–244.

Samé, A., F. Chamroukhi, G. Govaert, and P. Aknin. 2011. "Model-Based Clustering and Segmentation of Time Series with Changes in Regime." *Advances in Data Analysis and Classification* 5 (4): 301–321.

Shamshad, A., M. A. Bawadi, W. M. A. W. Hussin, T. A. Majid, and S. A. M. Sanusi. 2005. "First and Second Order Markov Chain Models for Synthetic Generation of Wind Speed Time Series." *Energy* 30 (5): 693–708.

Verbeek, J. J., N. Vlassis, and B. Kröse. 2003. "Efficient Greedy Learning of Gaussian Mixture Models." *Neural Computation* 15 (2): 469–485.

Viterbi, A. J. 1967. "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm." *IEEE Transactions on Information Theory* 13 (2): 260–269.

Zheng, K., Y. Li, and W. Xu. 2019. "Regime Switching Model Estimation: Spectral Clustering Hidden Markov Model." *Annals of Operations Research*, http://dx.doi.org/10.1007/s10479-019-03140-2.

Zhu, P., D. A. E. Acar, N. Feng, P. Jain, and V. Saligrama. 2019. "Cost Aware Inference for IoT Devices." *The 22nd International Conference on Artificial Intelligence and Statistics*, 89: 2770–2779.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Dynamic Allocation or Diversification: *A Regime-Based Approach to Multiple Assets*

Peter Nystrup, Bo William Hansen, Henrik Olejasz Larsen, Henrik Madsen, and Erik Lindström

The Journal of Portfolio Management

https://jpm.pm-research.com/content/44/2/62

ABSTRACT: *This article investigates whether regime-based asset allocation can effectively respond to changes in financial regimes at the portfolio level in an effort to provide better long-term results when compared to a static 60/40 benchmark. The potential benefit from taking large positions in a few assets at a time comes at the cost of reduced diversification. The authors analyze this trade-off in a multi-asset universe with great potential for static diversification. The regime-based approach is centered around a regime-switching model with time-varying parameters that can match financial markets' behavior and a new, more intuitive way of inferring the hidden market regimes. The empirical results show that regime-based asset allocation is profitable, even when compared to a diversified benchmark portfolio. The results are robust because they are based on available market data with no assumptions about forecasting skills.*

Dynamic Replication and Hedging: *A Reinforcement Learning Approach*

Petter N. Kolm and Gordon Ritter

The Journal of Financial Data Science

https://jfds.pm-research.com/content/1/1/159

ABSTRACT: *The authors of this article address the problem of how to optimally hedge an options book in a practical setting, where trading decisions are discrete and trading costs can be nonlinear and difficult to model. Based on reinforcement learning (RL), a well-established machine learning technique, the authors propose a model that is flexible, accurate and very promising for real-world applications. A key strength of the RL approach is that it does not make any assumptions about the form of trading cost. RL learns the minimum variance hedge subject to whatever transaction cost function one provides. All that it needs is a good simulator, in which transaction costs and options prices are simulated accurately.*

Regime-Based Versus Static Asset Allocation: *Letting the Data Speak*

Peter Nystrup, Bo William Hansen, Henrik Madsen, and Erik Lindström

The Journal of Portfolio Management

https://jpm.pm-research.com/content/42/1/103

ABSTRACT: *Regime shifts present a big challenge to traditional strategic asset allocation. This article investigates whether regime based asset allocation can effectively respond to changes in financial regimes at the portfolio level, in an effort to provide better long-term results than more static approaches can offer. The authors center their regime-based approach around a regime-switching model with time-varying parameters that can match financial markets' tendency to change behavior abruptly and the fact that the new behavior often persists for several periods after a change. In an asset universe consisting of a global stock index and a global government bond index, they show that, even without any level of forecasting skill, holding a static portfolio may not be optimal.*

Hyperparameter Optimization for Portfolio Selection

PETER NYSTRUP, ERIK LINDSTRÖM, AND HENRIK MADSEN

PETER NYSTRUP
is a postdoctoral fellow in the Centre for Mathematical Sciences at Lund University in Lund, Sweden, and in the Department of Applied Mathematics and Computer Science at the Technical University of Denmark in Lyngby, Denmark.
peter.nystrup@matstat.lu.se

ERIK LINDSTRÖM
is a professor and head of the Section for Mathematical Statistics in the Centre for Mathematical Sciences at Lund University in Lund, Sweden.
erik.lindstrom@matstat.lu.se

HENRIK MADSEN
is a professor and head of the Section for Dynamical Systems in the Department of Applied Mathematics and Computer Science at the Technical University of Denmark in Lyngby, Denmark.
hmad@dtu.dk

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

KEY FINDINGS

- The growing number of applications of machine-learning approaches to portfolio selection means that hyperparameter optimization becomes increasingly important. We propose a systematic approach to hyperparameter optimization by leveraging recent advances in automated machine learning and multiobjective optimization.
- We establish a connection between forecast uncertainty and holding- and trading-cost parameters in portfolio selection. We argue that they should be considered regularization parameters that can be adjusted in training to achieve optimal performance when tested subject to realized costs.
- We show that multiobjective optimization can find solutions with better risk–return trade-offs than manual, grid, and random search over hyperparameters for portfolio selection. At the same time, the solutions are more stable across in-sample training and out-of-sample testing.

ABSTRACT: *Portfolio selection involves a trade-off between maximizing expected return and minimizing risk. In practice, useful formulations also include various costs and constraints that regularize the problem and reduce the risk due to estimation errors, resulting in solutions that depend on a number of hyperparameters. As the number of hyperparameters grows, selecting their value becomes increasingly important and difficult. In this article, the authors propose a systematic approach to hyperparameter optimization by leveraging recent advances in automated machine learning and multiobjective optimization. They optimize hyperparameters on a train set to yield the best result subject to market-determined realized costs. In applications to single- and multiperiod portfolio selection, they show that sequential hyperparameter optimization finds solutions with better risk–return trade-offs than manual, grid, and random search over hyperparameters using fewer function evaluations. At the same time, the solutions found are more stable from in-sample training to out-of-sample testing, suggesting they are less likely to be extremities that randomly happened to yield good performance in training.*

TOPICS: ***Portfolio theory, portfolio construction, big data/machine learning****

Since Markowitz (1952) formulated the portfolio selection problem as a *trade-off* between maximizing expected return and minimizing risk, this has been the default approach in finance (Kolm, Tütüncü, and Fabozzi 2014). Building financial models from the ground up that model many or all aspects of the portfolio selection

problem is challenging. Specifically, each optimal submodel (e.g., transaction costs, holding costs, risk, return forecasts) is noisy and has its own approximation errors. We cannot expect that combining several noisy models together will be optimal. An alternative is to choose hyperparameters for each submodel based on other criteria, such as that of minimizing uncertainty or maximizing some performance metric (Ban, Karoui, and Lim 2018).

Multiperiod investment problems taking into account the stochastic nature of financial markets are usually solved in practice by scenario approximations of stochastic programming models, which are computationally challenging (see, e.g., Dantzig and Infanger 1993; Mulvey and Shetty 2004; Gülpınar and Rustem 2007; Pınar 2007). Recent literature has explored simulation-and-regression approaches to approximate the optimal policy through a large number of repetitions using *reinforcement learning* (Denault, Delage, and Simonato 2017; Zhang et al. 2018; Kolm and Ritter 2019). Both approaches require a good simulator.

Over the last decade, there has been an increasing interest in control-based approaches to solving multiperiod versions of the problem. This includes the use of *model predictive control* (MPC) for option hedging (Meindl and Primbs 2008; Bemporad, Bellucci, and Gabbriellini 2014), trade planning (Boyd et al. 2017), and portfolio selection (Herzog, Dondi, and Geering 2007; Nystrup et al. 2019). The idea is to control a portfolio based on forecasts of asset returns and relevant parameters. It is an intuitive approach with potential in practical applications because it is computationally fast. This makes it feasible to consider large numbers of assets and implement important constraints and costs.

The MPC formulations include hyperparameters such as risk, trading, holding, and shorting aversions that can be adjusted to achieve good performance in backtests and stress tests. Similarly, several studies have considered applications of regularization techniques to portfolio selection in the Markowitz (1952) framework (see, e.g., Clempner and Poznyak 2018; Aldridge 2019; Nystrup et al. 2019; and references therein) to address estimation errors that can significantly affect portfolio weights and result in poor out-of-sample performance. Including or amplifying transaction and holding costs and constraining portfolio weights are ways to regularize the optimization problem and reduce the risk due to estimation errors. For example, it can be beneficial to include a holding cost for short positions or a long-only constraint to limit or prevent short positions, even if the actual portfolio is not subject to such costs or constraints (Jagannathan and Ma 2003). In addition to the parameters in the portfolio selection problem, the set of hyperparameters could include hyperparameters in the forecasting model, such as memory and regularization parameters. We will leave the latter for future research to explore.

When the portfolio selection problem is convex and of the dimensions we consider in this article, it can be solved quickly using a custom solver. These solvers are fast enough to run in real time. The practical advantage of the high speed is the ability to carry out a large number of backtests quickly. Despite computational advances, it generally is not feasible to use brute force to find good values of the hyperparameters. Therefore, it is of interest to explore hyperparameter optimization in relation to portfolio selection.

We consider an innovative application of *multiobjective optimization* (MOO) to hyperparameter optimization and portfolio selection. We leverage recent advances in automated machine learning and apply them to problems of practical relevance in finance and portfolio selection. Through a combination of simulation and practical examples, we promote the idea of considering transaction and holding costs and constraints as surrogates or regularization parameters that can be adjusted in training to achieve optimal performance when tested subject to realized costs. Surprisingly, despite the work of Jagannathan and Ma (2003) alluding to this nearly two decades ago, the idea is not yet widespread within the field of finance.

Our work shows the potential value of hyperparameter optimization for both single- and multiperiod portfolio selection problems. In the single-period example, it is possible using MOO to find better risk–return trade-offs in less than half of the time required to use grid search. Moreover, to illustrate that transaction costs can reduce the risk due to estimation errors, we show empirically that when the signal-to-noise ratio of return forecasts deteriorates, the optimal level of transaction costs imposed in portfolio selection increases to prevent excessive noise trading. In the multiperiod example, we compare the result of MOO to random search. Using the same number of function evaluations, MOO finds better risk–return trade-offs on training data. To examine whether hyperparameter optimization

increases the risk of overfitting, we evaluate the results on test data and find that the optimized efficient frontier is more stable than the random-search frontier.

We begin with a brief introduction to hyperparameters and MOO. We then consider a single-period portfolio selection example from Boyd et al. (2017, Section 7.3). Subsequently, we consider an application of MPC to multiperiod portfolio selection based on the work of Nystrup et al. (2019). Finally, we conclude and provide suggestions for future work.

HYPERPARAMETER OPTIMIZATION

Every machine learning system has hyperparameters. An important task in automated machine learning is to automatically set these hyperparameters to optimize performance. Recent interest in complex and computationally expensive machine learning models with many hyperparameters, such as deep neural networks, has resulted in a resurgence of research on hyperparameter optimization (Feurer and Hutter 2019). Manual and grid search are the most widely used strategies, despite decades of research into global optimization (e.g., Nelder and Mead 1965; Broyden 1970; Powell 1994) and the publishing of several specialized hyperparameter optimization algorithms (e.g., Bergstra et al. 2011; Hutter, Hoos, and Leyton-Brown 2011; Snoek, Larochelle, and Adams 2012; Loshchilov, Glasmachers, and Beyer 2019).

Manual search is the simplest approach and gives researchers some degree of insight into the optimization problem, but it often fails to find the optimum. Grid search and random search are extensions that are simple to implement and parallelize. *Grid search* is done by specifying a finite set of values for each hyperparameter and evaluating the Cartesian product of these sets. It provides a systematic approach and typically finds a better solution than purely manual sequential optimization (in the same amount of time), but it scales poorly with the number of hyperparameters[1] and involves fitting the full model for values of hyperparameters that are very unpromising.

Random search has been proven to yield a faster exploration of the hyperparameter space than grid search, especially when some hyperparameters are much more important than others, which is often the case

[1] The required number of function evaluations grows exponentially with the dimension of the configuration space.

EXHIBIT 1
Black-Box Scenario

(Bergstra and Bengio 2012). It is a useful method for initializing the search process because it explores the entire configuration space and thus often finds settings with reasonable performance. Random search is a useful baseline because it makes no assumptions about the problem being optimized and, given enough resources, achieves performance arbitrarily close to the optimum. However, none of these methods make use of previous evaluations to make an informed decision for the next iterate. As such, convergence to a global optima can be very slow (Pedregosa 2016).

In general, every *black-box* optimization method can be applied to hyperparameter optimization. In a black-box scenario, as illustrated in Exhibit 1, knowledge about an objective function $f\colon X \to \mathbb{R}$, to be optimized on some space X, is restricted to the handling of a device (e.g., simulation code) that delivers the value of $f(x)$ for any input $x \in X$.

In our case, the output values are return and risk in a backtest for given values of a set of hyperparameters. Running a backtest, say with daily trading over a five-year period, is often computationally demanding, meaning that each function evaluation is expensive. Gradient-based methods are not applicable because this is a black-box scenario. Rather, we consider a state-of-the-art *evolution strategy*, which, similar to Bayesian optimization, uses past function evaluations to update the distribution from which new candidate solutions are sampled (Voß, Hansen, and Igel 2010). In addition, the update of the distribution parameters follows a natural gradient approximation of the expected objective-function value (Hansen and Ostermeier 2001).

The purpose of automated hyperparameter optimization is to reduce the human effort necessary for applying machine learning, improve the performance of machine learning algorithms, and improve the reproducibility and fairness of scientific studies (Feurer and Hutter 2019). Automated hyperparameter optimization is clearly more reproducible than

manual search. In particular, different solutions can only be compared fairly if they all receive the same level of tuning for the problem at hand. The best strategy for how to avoid *overfitting* varies across hyperparameter optimization problems and remains an open problem. Automated approaches are a step in the right direction. As such, the objective of automated hyperparameter optimization extends to financial applications.

MULTIOBJECTIVE OPTIMIZATION

In practical applications it is often necessary to trade off two or more objectives. For example, the Markowitz (1952) problem of maximizing return and minimizing risk is a MOO problem. A MOO problem can be formulated as

$$\begin{aligned} &\text{minimize} && (f_1(x), f_2(x), \ldots, f_K(x)) \\ &\text{subject to} && x \in X, \end{aligned} \tag{1}$$

where $K \geq 2$ is the number of objectives, and X is the feasible set of decision vectors. The feasible set is typically defined by some constraint functions. If some objective function is to be maximized, it is equivalent to minimize its negative.

Efficient Frontier

In MOO, one feasible solution $x \in X$ that optimizes all objective functions simultaneously typically does not exist. Therefore, attention is paid to Pareto-optimal solutions (i.e., solutions that cannot be improved in any of the objectives without degrading at least one of the other objectives). The set of Pareto-optimal outcomes is often called the Pareto front or *efficient frontier.*

The goal of MOO is to find a diverse set of Pareto-optimal solutions, which provide insights into the trade-offs between the objectives. When approaching a MOO problem by linearly aggregating all objectives into a scalar function, as is customarily done in financial applications (e.g., Zenios 2007, Chapter 3), each weighting of the objectives yields only a subset of Pareto-optimal solutions (usually only a single solution). Hence, various optimization trials with different aggregations become necessary to find a diverse set of Pareto-optimal solutions (Igel, Hansen, and Roth 2007). For example, the standard approach in portfolio selection is to solve the Markowitz (1952) problem:

$$\begin{aligned} &\text{maximize} && \mu^T w - \gamma^{\text{risk}} w^T \Sigma w \\ &\text{subject to} && \mathbf{1}^T w = 1, \end{aligned} \tag{2}$$

where the variable w is the unknown portfolio weights, μ is the vector of expected returns, Σ is the covariance matrix for the returns, and $\mathbf{1}$ is a column vector with all entries being one, for several different values of the risk-aversion parameter γ^{risk} to trace the efficient frontier.

MULTIOBJECTIVE COVARIANCE-MATRIX ADAPTATION EVOLUTION STRATEGY

We use the multiobjective covariance-matrix adaptation evolution strategy (MO–CMA–ES) from Voß, Hansen, and Igel (2010) that is implemented in the Distributed Evolutionary Algorithms in Python (DEAP) framework by Fortin et al. (2012). It is a stochastic, derivative-free method for numerical optimization of difficult (nonconvex, ill-conditioned, multimodal, rugged, noisy) continuous optimization problems. We use the MO–CMA–ES because it is one of the best known and most competitive black-box optimization algorithms (Feurer and Hutter 2019).

The MO–CMA–ES belongs to the class of evolutionary algorithms. In each generation (iteration), new individuals (candidate solutions, denoted as x) are generated by varying the current parental individuals. Some individuals then are selected to become the parents in the next generation based on their objective value $f(x)$. Over the generation sequence, individuals with better and better f-values are selected.

The way that past functions evaluations are used to update the parameters of the distribution from which new candidate solutions are sampled is different from Bayesian optimization. In an evolution strategy, new candidate solutions are sampled according to a multivariate Gaussian distribution. Dependencies between the variables are represented by a covariance matrix. Updates of the mean and covariance matrix are done by maximizing the likelihood of previously successful candidate solutions and search steps, respectively.

Adaptation of the covariance matrix amounts to learning a second-order model of the underlying objective function, similar to the approximation of the inverse

Hessian matrix in the quasi-Newton method in classical optimization (e.g., Broyden 1970; Wills and Schön 2017). In contrast to most classical methods, fewer assumptions on the nature of the underlying objective function are made. Only the ranking among candidate solutions is exploited for learning the sample distribution, and neither derivatives nor the function values themselves are required by the method (Hansen and Ostermeier 2001).

SINGLE-PERIOD PORTFOLIO SELECTION

In this section, we consider the single-period portfolio selection example from Boyd et al. (2017, Section 7.3).

Problem Formulation

Given a current portfolio of N risky assets and one risk-free asset with weights $w_t \in \mathbb{R}^{N+1}$, the single-period portfolio selection problem is to

$$\begin{aligned} \text{maximize} \quad & \hat{\mu}_{t+1|t}^T w_{t+1} - \gamma^{\text{trade}} \hat{\phi}_{t+1|t}^{\text{trade}}(w_{t+1} - w_t) - \gamma^{\text{hold}} \hat{\phi}_{t+1|t}^{\text{hold}}(w_{t+1}) \\ & - \gamma^{\text{risk}} w_{t+1}^T \hat{\Sigma}_{t+1|t} w_{t+1} \\ \text{subject to} \quad & \mathbf{1}^T w_{t+1} = 1, \end{aligned} \tag{3}$$

with variable $w_{t+1} \in \mathbb{R}^{N+1}$, where $\hat{\mu}_{t+1|t}$ and $\hat{\Sigma}_{t+1|t}$ are the forecasts made at time t of the returns and covariances at time $t + 1$. In addition, a maximum leverage constraint $||(w_{t+1})_{1:N}||_1 \leq L^{\max}$ is imposed with $L^{\max} = 3$, where $(w_{t+1})_{1:N} = ((w_{t+1})_1, \ldots, (w_{t+1})_N)$ is the vector of risky-asset weights.

Transaction and holding costs. In Equation 3, $\hat{\phi}^{\text{trade}}$ and $\hat{\phi}^{\text{hold}}$ are estimates of the actual, realized transaction- and holding-cost functions, although they could be arbitrary (convex) functions found to give good performance in backtesting. The holding cost includes a charge of 1 bp for all risky assets, and the transaction cost is volume and volatility dependent with an assumed bid–ask spread of 5 bps. We refer to Boyd et al. (2017, Section 7.3) for the specific details of the example.

Hyperparameters. Equation 3 includes three hyperparameters that need to be selected:

1. A risk-aversion parameter $\gamma^{\text{risk}} \geq 0$ that discourages holding portfolios with high variance. Alternatively, the expected return can be maximized subject to an upper limit on the portfolio variance, or the portfolio variance can be minimized subject to a lower limit on the expected return (Kolm, Tütüncü, and Fabozzi 2014).
2. A trading-aversion factor $\gamma^{\text{trade}} \geq 0$ that scales the transaction-cost function. Increasing the factor will deter trading or reduce turnover; decreasing it will increase trading and turnover. It is even possible to think of $1/\gamma^{\text{trade}}$ as the number of periods over which transaction costs are amortized (Grinold 2006). Keeping $\gamma^{\text{trade}} = 1$ would mean that we require trades to be profitable after a single period; otherwise, we will not trade.
3. A holding-aversion parameter $\gamma^{\text{hold}} \geq 0$ that scales the holding-cost function to encourage, or discourage, holding positions that incur holding costs, such as short positions.

Simulation Study

We follow Boyd et al. (2017) by considering an asset universe consisting of $N = 430$ stocks from the S&P 500 Index with each backtest consisting of five years of daily trading from January 2012 through December 2016.

Forecasts

Return forecasts are constructed by adding zero-mean noise to realized returns of individual stocks from the S&P 500 Index and then rescaling to obtain return estimates that would (approximately) minimize mean squared error.

For all t, the return estimates for risky assets are

$$(\hat{\mu}_t)_{1:N} = \alpha((r_t)_{1:N} + \epsilon_t), \tag{4}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$ are independent. We follow Boyd et al. (2017) and use noise variance $\sigma_\epsilon^2 = 0.02$, so the noise components have a standard deviation of around 14%, approximately a factor of 10 larger than the standard deviation of the realized returns. The scale factor α is chosen to minimize the mean squared error $\mathrm{E}[((\hat{\mu}_t)_{1:N} - (r_t)_{1:N})^2]$, if we think of r_t as a random variable with variance σ_r, that is, $\alpha = \sigma_r^2/(\sigma_r^2 + \sigma_\epsilon^2)$. We use the typical value $\sigma_r^2 = 0.0005$ (i.e., a realized return standard deviation of around 2%) so $\alpha = 0.024$. The typical return forecast is on the order of ±0.3%. This corresponds to an information ratio $\sqrt{\alpha} \approx 0.15$, which

is on the high end of what might be expected in practice (Grinold and Kahn 2000).

Boyd et al. (2017) argued that with this level of noise and scaling, the return forecasts have an accuracy on the order of what might be expected from a proprietary forecast. For example, across all the assets and all days, the sign of the predicted return agrees with the sign of the real return around 54% of the time. Later we will explore what happens when the noise variance σ_ϵ^2 is doubled to 0.04, but for now we will stick with these values.

Total market volumes and daily volatilities (used in the transaction cost model) are estimated as moving averages of the realized values with a window of length 10. Finally, the covariance forecasts are constructed using a factor model estimated on a two-year rolling window of past realized returns. Once again, we refer to Boyd et al. (2017, Section 7.3) for specific details on the example.

Results

After a couple of coarse grid searches to narrow the region of interest, Boyd et al. (2017) performed a detailed hyperparameter search testing all combinations of

$$\begin{aligned}\gamma^{\text{risk}} &= 0.1, 0.178, 0.316, 0.562, 1, 2, 3, 6, 10, 18, 32, 56, 100,\\ &\quad 178, 316, 562, 1000,\\ \gamma^{\text{trade}} &= 5.5, 6, 6.5, 7, 7.5, 8,\\ \gamma^{\text{hold}} &= 0.1, 1, 10, 100, 1000,\end{aligned}$$

for a total of 510 backtest simulations. This takes about 12 hours on a standard laptop PC with a 1.9 GHz Intel i7 processor using its eight cores in parallel.

For comparison, we perform hyperparameter optimization by maximizing the return and minimizing the standard deviation using the MO–CMA–ES with initial step size $\sigma_0 = 0.1$. We map these ranges for the hyperparameters to the unit interval [0, 1] using power transformations. For example, $\gamma^{\text{risk}} = 10^{-1+4\theta_1}$, where θ_1 is the transformed risk-aversion parameter. Each generation consists of 20 individuals, and the first generation is sampled randomly from the three-dimensional unit cube.

Exhibit 2 compares the efficient frontier based on grid search to the results of MOO using the MO–CMA–ES. With only 240 function evaluations compared with 510 using grid search, it is possible to find better risk–return trade-offs using the MO–CMA–ES in a fraction of time. The computations involved in considering previous evaluations when sampling a new generation are negligible compared with the time it takes to run a five-year backtest. The objective values for an entire generation of hyperparameter values can be computed in parallel, similar to grid search. Hence, the overhead compared with grid search is minimal, and it makes sense to compare the methods in terms of the number of backtest evaluations.

Holding aversion. As noted by Boyd et al. (2017), we find that along the efficient frontier the optimal values of the holding-cost multiplier γ^{hold} (not shown) are much larger than one. This is an example of regularization in single-period portfolio selection. The large holding-aversion parameter tells the portfolio selection to avoid short positions, even though the actual holding cost (used by the simulator to update the portfolio each day) is only 1 bp, and the result is overall better portfolio performance. This is consistent with the finding by Jagannathan and Ma (2003) that constraining portfolio weights to be nonnegative can reduce the risk in estimated optimal portfolios even when the constraints are wrong.

Different Noise Levels

To further illustrate the benefits of hyperparameter optimization for single-period portfolio selection, we repeat the simulation study, but this time we double the variance of the noise added to the realized returns to construct return forecasts from $\sigma_\epsilon^2 = 0.02$ to $\sigma_\epsilon^2 = 0.04$. This corresponds to an information ratio $\sqrt{\alpha} \approx 0.1$ (i.e., by increasing the noise variance, the signal-to-noise ratio worsens).

Exhibit 3, Panel A, shows the efficient frontiers based on MOO using 240 evaluations for the two levels of noise in the return forecasts. The frontier for $\sigma_\epsilon^2 = 0.02$ is the same as in Exhibit 2. As expected, the weaker signal-to-noise ratio pushes the efficient frontier downward. With the added noise, the Sharpe ratios of the portfolios along the efficient frontier are closer to one than to two.

Exhibit 3, Panel B, shows the optimal values of the trading-aversion parameter γ^{trade} for the points on the efficient frontiers. Although the optimal value varies with the risk level, it is clear that optimal values of the parameter are higher when the forecasts are noisier. Roughly speaking, when the noise variance $\sigma_\epsilon^2 = 0.02$,

EXHIBIT 2

Efficient Frontiers Based on Grid Search and Multiobjective Optimization

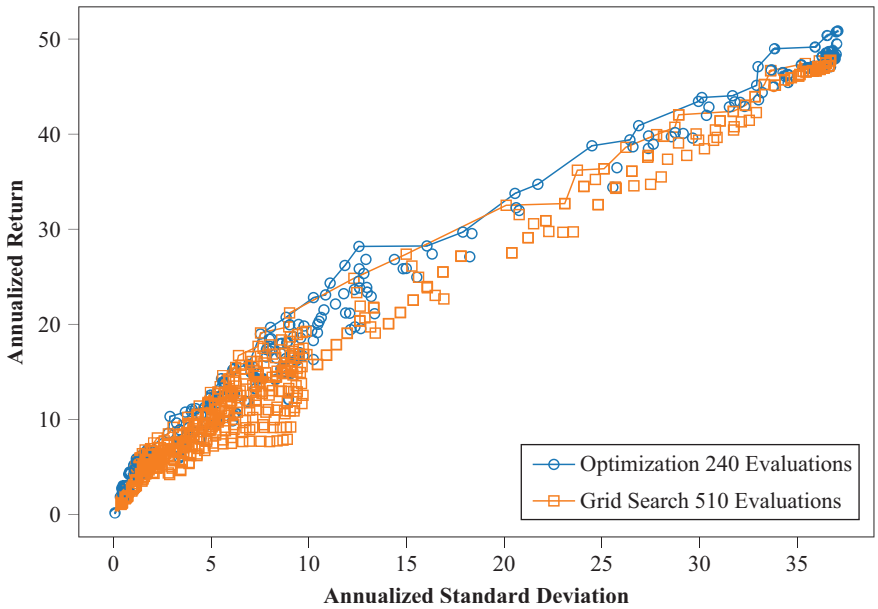
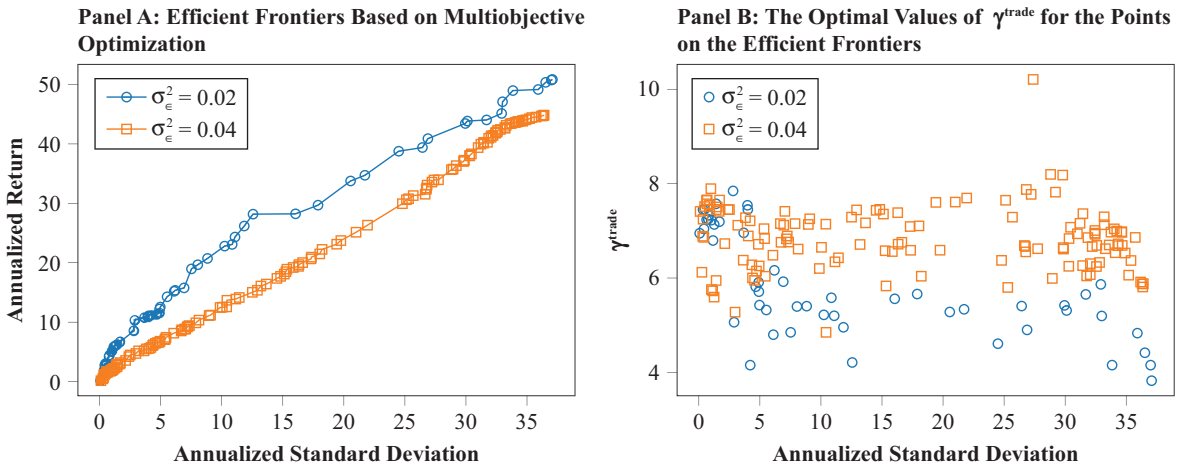


EXHIBIT 3

Efficient Frontiers Based on Multiobjective Optimization for Different Levels of Noise in the Return Forecasts



Note: The results are based on 240 function evaluations.

the optimal values of γ^{trade} are around five, compared with around seven when $\sigma_\epsilon^2 = 0.04$. In other words, when the signal-to-noise ratio of the return forecasts deteriorates, the optimal level of the transaction-cost multiplier increases to prevent excessive trading, despite realized transaction costs remaining unchanged.

The optimization of the hyperparameters for two different levels of noise clearly shows the connection between the trading-aversion parameter and return forecast error. Similar to the holding-aversion parameter discussed earlier, the trading-aversion parameter regularizes the portfolio selection problem and reduces the

risk due to estimation error. This would have been much harder to show without the use of MOO. Using grid search would have required significantly more function evaluations, and at the same time we would have had to run extra checks to ensure the grid was centered around the range of values of interest of the hyperparameters to see the differences. With a misaligned or too coarse grid, we might not have been able to detect the differences.

It is both a strength and weakness that the hyperparameters in portfolio selection have clear economic interpretations in terms of costs and constraints. It makes them intuitively easy to understand and justify, but it is easy to confuse them with their realized counterparts. We argue that the hyperparameters should be considered regularization parameters that can be adjusted in training to achieve optimal performance when tested, subject to realized costs. As illustrated in this section through realistic simulation examples, it is rarely optimal to keep the transaction- and holding-cost aversion parameters equal to one (i.e., to use the best estimate of the realized costs in the portfolio selection). In the next section, we take this idea even further and consider trading- and holding-regularization functions that do not have to be a multiple of the estimated realized cost functions.

MULTIPERIOD PORTFOLIO SELECTION

We continue by considering an application of MPC to multiperiod portfolio selection based on the work of Nystrup et al. (2019), where return and covariance forecasts are generated using an actual forecasting model. In a multiperiod framework, we can naturally handle time-varying forecasts, market impact costs, and other phenomena that require planning multiple steps ahead (Boyd et al. 2017). To examine whether hyperparameter optimization increases the risk of overfitting, we compare results using both training and test data.

Model Predictive Control

MPC is based on the simple idea that, to determine the trades to make, all future (unknown) quantities are replaced by their forecasted values over a planning horizon H. For example, future returns are replaced by their expected values $\hat{\mu}_{\tau|t}$, $\tau = t+1, \ldots, t+H$, where $\hat{\mu}_{\tau|t}$ is the forecast made at time t of the return at time τ. This leads to a deterministic optimization problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{\tau=t+1}^{t+H} (\hat{\mu}_{\tau|t}^T w_\tau - \hat{\phi}_{\tau|t}^{\text{trade}}(w_\tau - w_{\tau-1}) \\ & \qquad - \hat{\phi}_{\tau|t}^{\text{hold}}(w_\tau) - \gamma^{\text{risk}} w_\tau^T \hat{\Sigma}_{\tau|t} w_\tau) \\ \text{subject to} \quad & \mathbf{1}^T w_\tau = 1, \quad \tau = t+1, \ldots, t+H, \end{aligned} \tag{5}$$

with variables $w_{t+1}, \ldots, w_{t+H} \in \mathbb{R}^{N+1}$. Note that w_t is not a variable but rather the known, current portfolio weights.

Solving the problem in Equation 5 yields an optimal sequence of weights $w_{t+1}^*, \ldots, w_{t+H}^*$. The difference of this sequence is a plan for future trades over the planning horizon $H = 15$ under the unrealistic assumption that all future (unknown) quantities will be equal to their forecasted values. Only the first trade $w_{t+1}^* - w_t$ in the planned sequence of trades is executed. In the next step, the process is repeated, starting from the new portfolio w_{t+1}. The planning horizon can typically be much shorter than the investment horizon without affecting the solution (Boyd et al. 2017). MPC for stochastic systems is a suboptimal control strategy; however, it uses new information advantageously and is better than pure open-loop control.

Constraints

We consider long-only portfolios with maximum holdings of 40% per risky asset; that is, $\mathbf{0} \leq (w_\tau)_{1:N} \leq (0.4)\,\mathbf{1}$, $\tau = t+1, \ldots, t+H$. Imposing constraints on the portfolio weights is a way to improve out-of-sample performance. It is equivalent to shrinking the covariance matrix (Jagannathan and Ma 2003; Ledoit and Wolf 2003, 2004; DeMiguel et al. 2009; Li 2015). Different constraints correspond to different prior beliefs about the asset weights.

In addition, the portfolios are subject to a maximum leverage constraint $||(w_{t+1})_{1:N}||_1 \leq L^{\max}$ with $L^{\max} = 2$. We assume that the only cost associated with leveraging a portfolio is the risk-free rate. We follow Nystrup et al. (2019) in specifying these holding and leverage constraints, although we note that the minimum and maximum holdings as well as the leverage limit could be considered hyperparameters to be selected along with the risk-aversion and cost parameters discussed next. This would render grid search infeasible for all practical purposes.

Transaction and Holding Penalties

As in the single-period setting, $\hat{\phi}^{\text{trade}}$ and $\hat{\phi}^{\text{hold}}$ can be estimates of actual transaction- and holding-cost functions or arbitrary (convex) regularization functions found to give good performance in backtesting. This time we consider the latter approach. That is, we explore different trading- and holding-regularization functions while assuming a realized transaction cost of 10 bps per one-way transaction and no realized holding cost when evaluating performance in backtests.

To regularize the portfolio selection problem and reduce the risk of trading too much, Nystrup et al. (2019) included a penalty for trading:

$$\phi_t^{\text{trade}}(w_t - w_{t-1}) = \kappa_1 \left\| w_t - w_{t-1} \right\|_1 + \kappa_2 \left\| w_t - w_{t-1} \right\|_2^2, \quad (6)$$

in the objective function, where κ_1 and κ_2 are penalty factors. This could reflect actual transaction costs or a conservatism toward trading (e.g., due to the uncertainty related to the parameter estimates and forecasts).

The weighted elastic-net trading penalty (Equation 6) is a convex combination of ℓ_1- and squared ℓ_2-norm penalties. It reduces the number of trades like the ℓ_1 penalty and the size of trades like the squared ℓ_2 penalty. The ℓ_1 penalty is similar to the standard proportional transaction cost (which we assume when evaluating performance) and is a convex relaxation of constraints on the number of trades. The squared ℓ_2 penalty is used to model price impact (Almgren and Chriss 2001, Boyd et al. 2017); it shrinks together trades in correlated assets and splits trades over multiple days.[2]

Furthermore, Nystrup et al. (2019) included a holding penalty:

$$\phi_t^{\text{hold}}(w_t) = \rho_1 \left\| w_t \right\|_1 + \rho_2 \left\| w_t \right\|_2^2, \quad (7)$$

where ρ_1 and ρ_2 are penalty factors. Including holding costs is a means to control portfolio leverage, as discussed in the previous section. When considering long–short portfolios, this could include an additional holding penalty specific to short positions, which would be another hyperparameter.

The weighted elastic-net holding penalty (Equation 7) can be justified by reformulating the mean–variance criterion as a robust optimization problem (Ho, Sun, and Xin 2015; Boyd et al. 2017). It reduces the number of holdings like the ℓ_1 penalty and the size of holdings like the squared ℓ_2 penalty. The ℓ_1 penalty is a convex relaxation of constraints on the number of holdings. It can be regarded as a shrinkage estimator of the expected return (Stein 1956; Fabozzi, Huang, and Zhou 2010). The squared ℓ_2 penalty shrinks together holdings in correlated assets; it corresponds to adding a diagonal matrix to the forecasted covariance matrix in Equation 5, similar to a Stein-type shrinkage estimator (Ledoit and Wolf 2004). Even when a portfolio is not subject to any holding cost, it can still be beneficial to include holding penalties in the optimization problem.

Considering the specific parameters in the trading- and holding-regularization functions as hyperparameters is a flexible alternative to introducing trading- and holding-aversion parameters that scale the realized cost functions. In total, the hyperparameter set consists of a risk-aversion parameter γ^{risk}, trading-penalty factors κ_1 and κ_2, and holding-penalty factors ρ_1 and ρ_2. Similar to Nystrup et al. (2019), we assume that trading- and holding-penalty parameters are the same for all risky assets and zero for the risk-free asset. Without this assumption, the hyperparameter set would be considerably larger and grid search would be completely infeasible.

Data

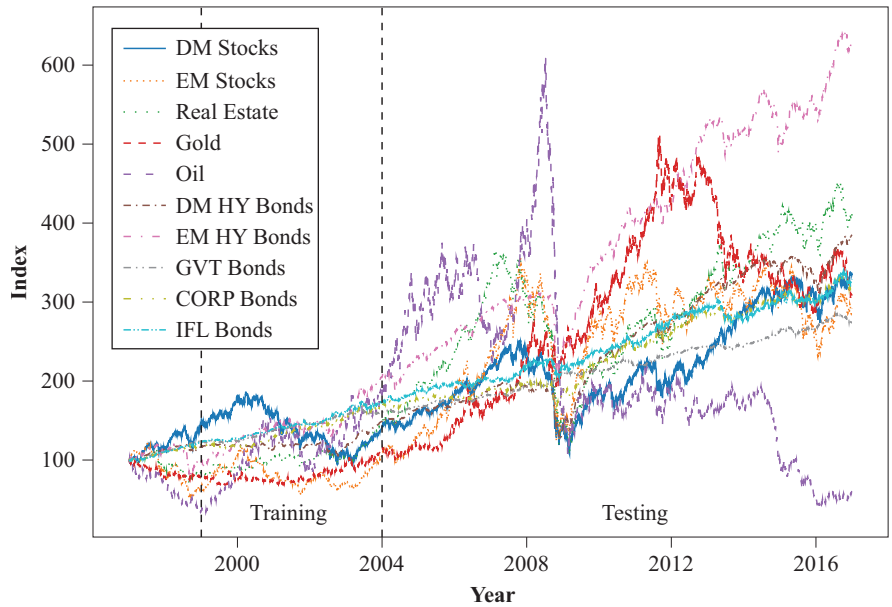
Exhibit 4 shows the daily index data considered by Nystrup et al. (2019). The 10 indexes are developing market (DM) and emerging market (EM) stocks, listed DM real estate, DM and EM high-yield bonds, gold, oil, corporate bonds, inflation-linked bonds, and government bonds. All indexes measure the net total return in US dollars with a total of 5,185 daily closing prices per index covering the period from 1997 through 2016.

We use the first two years for model initialization, the next five years for training, and the last 13 years for testing. As always when considering financial data, the split between training and test sets is pivotal. Most importantly, the five years from 1999 to 2003 include both bull and bear markets. We use a different split than Nystrup et al. (2019), who used data before 1999 for training.

We use their forecasts of μ and Σ, which are based on a multivariate, Gaussian hidden Markov model with two states estimated using a sequential version of the

[2] Price impact is the price movement against the trader that tends to occur when a large order is executed.

EXHIBIT 4
Daily Index Data Used for Training and Testing



Note: CORP = corporate; DM = developing markets; EM = emerging markets; GVT = government; HY = high yield; IFL = inflation linked.

Expectation–Maximization algorithm with a fading memory and without any forward looking. Every day when a new observation is recorded, the model estimates the probability of that observation belonging to the risk-on and risk-off states, respectively. Based on the estimated probabilities, the model parameters are updated, and subsequently μ and Σ are forecasted 15 days into the future starting from the estimated probability distribution over the current market state. Several other studies have considered asset allocation based on regimes inferred from financial returns using regime-switching models (Guidolin and Timmermann 2007; Bulla et al. 2011; Nystrup et al. 2015, 2017; Nystrup, Madsen, and Lindström 2018).

Results

In this section, we focus on comparing the results of MOO and random search (i.e., independent draws from a uniform density from the same configuration space as would be spanned by a regular grid). Bergstra and Bengio (2012) showed empirically and theoretically that randomly chosen trials are more efficient for hyperparameter optimization than trials on a grid. Thus, random search is a natural baseline against which to compare adaptive (sequential) hyperparameter optimization algorithms.

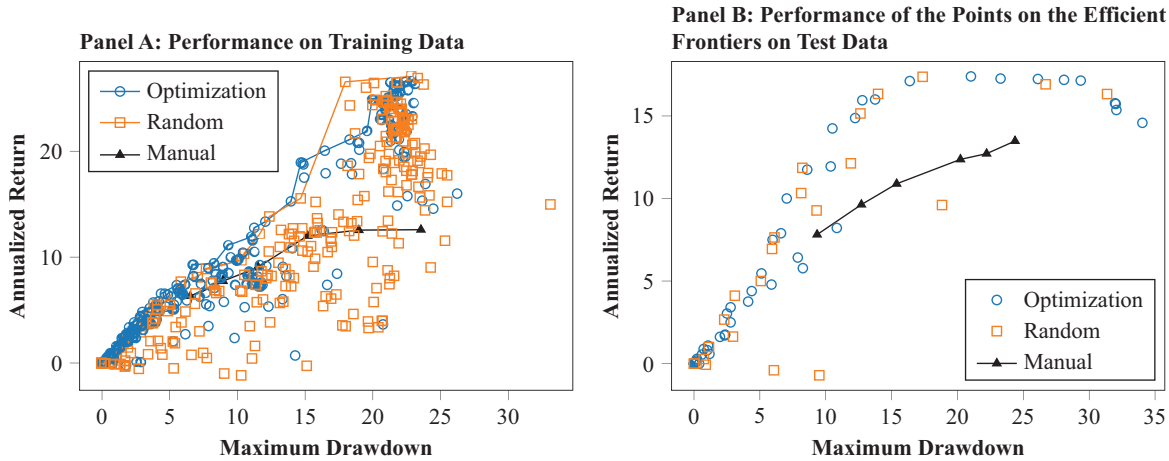
The portfolio selection problem (Equation 5) is solved using CVXPY (Diamond and Boyd 2016) with the open-source solver ECOS (Domahidi, Chu, and Boyd 2013). We perform hyperparameter optimization by maximizing the return and minimizing the maximum drawdown (MDD)[3] over the five-year training period using the MO–CMA–ES with initial step size $\sigma_0 = 0.1$. Note that minimizing MDD is a different objective than in Equation 5. As already discussed, it can be advantageous not to consider the same objective function as in the underlying model when optimizing hyperparameters.

As in the single-period setting, we map the regions of interest for the hyperparameters to the unit interval using power transformations: $\gamma^{\text{risk}} = 10^{-3+5\theta_1}$, $\kappa_1 = 10^{-7+6\theta_2}$, $\kappa_2 = 10^{-7+6\theta_3}$, $\rho_1 = 10^{-7+5\theta_4}$, and $\rho_2 = 10^{-7+5\theta_5}$. These intervals are selected based on initial experimentation and are wider than those considered by Nystrup et al. (2019).

[3] The maximum drawdown is the largest relative decline from a historical peak in the index value.

EXHIBIT 5

Efficient Frontiers Based on Random Search and Multiobjective Optimization Compared with Manual Search



Note: The results are based on 240 function evaluations.

Each generation consists of 20 individuals, and the first generation is sampled randomly from the five-dimensional unit cube.

Training

Exhibit 5, Panel A, compares the efficient frontiers based on random search and MOO using 240 evaluations on training data. The annualized return is in excess of the risk-free rate, which is the daily equivalent of the yield on a one-month US Treasury bill. The frontiers are compared with the frontier found by Nystrup et al. (2019), who did a coarse manual search in combination with linear aggregation for $\gamma^{\text{risk}} = 1, 3, 5, 10, 15, 25$. Both random search and MOO using the MO–CMA–ES find better risk–return trade-offs than the points found by Nystrup et al. (2019). The MO–CMA–ES yields slightly better results than random search for MDDs smaller than 15%. It is evident that the search points are much more scattered for random search compared with the MO–CMA–ES.

The excess return that can be achieved for the same level of MDD compared with the manual search solutions amounts to several percentage points. Although it is interesting to see just how large the difference can be in a realistic application based on real forecasts, the value of hyperparameter optimization depends on the particular application, the number of hyperparameters, computational constraints, forecast strength, and other factors.

Testing

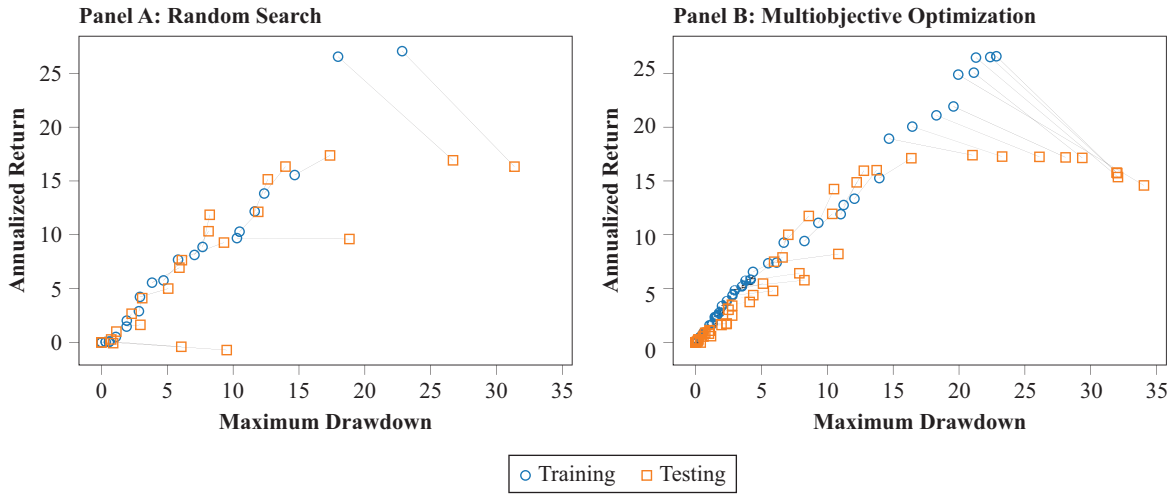
An important concern is the extent to which the excess performance is a result of overfitting and hence will not persist when tested out of sample. After all, it seems reasonable to assume that optimizing the hyperparameters increases the risk of overfitting. This presumption was why Nystrup et al. (2019) did not consider hyperparameter optimization in greater detail.

In Exhibit 5, Panel B, the points on the efficient frontiers in Exhibit 5, Panel A, are evaluated on the test data. Once again, random search and MOO using the MO–CMA–ES yield significantly better risk–return trade-offs than manual search. This difference refutes the argument that a coarse manual or grid search decreases the risk of overfitting and increases out-of-sample robustness. The majority of the points lie close together for random search and the MO–CMA–ES, with the important exception of a handful of points that deviate considerably in a negative direction for random search.

Exhibit 6 illustrates the connection between the points on the efficient frontiers based on training data and their performance on the test data. Most of the points lie fairly close together. Yet, for random search, a few of the points that had very low MDDs in sample experience substantially higher MDDs out of sample. For the MO–CMA–ES, only the points that had the highest MDDs on the training data realize substantially higher MDDs in testing. However, the mapping

EXHIBIT 6

Connection among the Points on the Efficient Frontier Based on Training Data and Their Performance on Test Data



is consistent and suggests that risk taking above MDDs of 15% is not rewarded in the test period.

Although in sample the efficient frontier based on MOO was only slightly better than that based on random search, we notice important differences out of sample. In particular, the mapping from training to testing of the points on the efficient frontier is less stable when based on random search, as is evident from Exhibit 6.

In summary, the excess performance achieved through more careful hyperparameter optimization seems to persist from the in-sample training to the out-of-sample testing. Naturally, this finding is specific to the particular case, data period, train–test split, and so on. Nevertheless, it is an important finding that shows the potential for adding value both in and out of sample through a systematic approach to hyperparameter optimization.

CONCLUSION

The growing number of applications of machine learning approaches to portfolio selection means that hyperparameter optimization becomes increasingly important. Through a number of examples, we illustrated the connection between forecast uncertainty and holding- and trading-regularization parameters in portfolio selection. In particular, we showed that when the signal-to-noise ratio of return forecasts deteriorates, the optimal level of the transaction-cost multiplier increases to prevent excessive trading, although realized transaction costs remain unchanged.

We showed that MOO can find solutions with better risk–return trade-offs than manual, grid, and random search over hyperparameters. At the same time, solutions found using MOO were more stable across in-sample training and out-of-sample testing, indicating that improvements achieved through hyperparameter optimization were not a result of overfitting.

The value of hyperparameter optimization depends on the particular application, the number of hyperparameters, computational constraints, forecast strength, and other factors. However, when using a sequential hyperparameter optimization algorithm, it is less likely that solutions found are extremities that randomly happened to yield good performance in training.

Future work involving hyperparameter optimization in relation to portfolio selection should consider MOO as an alternative to the more common methods of manual and grid search. In addition, as hyperparameter optimization becomes increasingly important in quantitative finance and financial data science, future work should explore systematic approaches to combat overfitting.

ACKNOWLEDGMENTS

This work was supported by Vergstiftelsen and the Centre for IT-Intelligent Energy Systems (CITIES) project funded in part by Innovation Fund Denmark under Grant No. 1305-00027B. The authors are indebted to Ilya Loshchilov for proposing this article.

REFERENCES

Aldridge, I. 2019. "Big Data in Portfolio Allocation: A New Approach to Successful Portfolio Optimization." *The Journal of Financial Data Science* 1 (1): 45–63.

Almgren, R., and N. Chriss. 2001. "Optimal Execution of Portfolio Transactions." *Journal of Risk* 3 (2): 5–39.

Ban, G. Y., N. E. Karoui, and A. E. B. Lim. 2018. "Machine Learning and Portfolio Optimization." *Management Science* 64 (3): 1136–1154.

Bemporad, A., L. Bellucci, and T. Gabbriellini. 2014. "Dynamic Option Hedging via Stochastic Model Predictive Control Based on Scenario Simulation." *Quantitative Finance* 14 (10): 1739–1751.

Bergstra, J., and Y. Bengio. 2012. "Random Search for Hyper-Parameter Optimization." *Journal of Machine Learning Research* 13: 281–305.

Bergstra, J. S., R. Bardenet, Y. Bengio, and B. Kégl. "Algorithms for Hyper-Parameter Optimization." In *Advances in Neural Information Processing Systems*, pp. 2546–2554. Red Hook, NY: Curran Associates. 2011.

Boyd, S., E. Busseti, S. Diamond, R. N. Kahn, K. Koh, P. Nystrup, and J. Speth. 2017. "Multi-Period Trading via Convex Optimization." *Foundations and Trends in Optimization* 3 (1): 1–76.

Broyden, C. G. 1970. "The Convergence of a Class of Double-Rank Minimization Algorithms 1. General Considerations." *IMA Journal of Applied Mathematics* 6 (1): 76–90.

Bulla, J., S. Mergner, I. Bulla, A. Sesboü, and C. Chesneau. 2011. "Markov-Switching Asset Allocation: Do Profitable Strategies Exist?" *Journal of Asset Management* 12 (5): 310–321.

Clempner, J. B., and A. S. Poznyak. 2018. "Sparse Mean–Variance Customer Markowitz Portfolio Optimization for Markov Chains: A Tikhonov's Regularization Penalty Approach." *Optimization and Engineering* 19 (2): 383–417.

Dantzig, G. B., and G. Infanger. 1993. "Multi-Stage Stochastic Linear Programs for Portfolio Optimization." *Annals of Operations Research* 45 (1): 59–76.

DeMiguel, V., L. Garlappi, F. Nogales, and R. Uppal. 2009. "A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms." *Management Science* 55 (5): 798–812.

Denault, M., E. Delage, and J. G. Simonato. 2017. "Dynamic Portfolio Choice: A Simulation-and-Regression Approach." *Optimization and Engineering* 18 (2): 369–406.

Diamond, S., and S. Boyd. 2016. "CVXPY: A Python-Embedded Modeling Language for Convex Optimization." *Journal of Machine Learning Research* 17 (83): 1–5.

Domahidi, A., E. Chu, and S. Boyd. "ECOS: An SOCP Solver for Embedded Systems." In *Proceedings of the 12th European Control Conference*, pp. 3071–3076. New York: IEEE. 2013.

Fabozzi, F. J., D. Huang, and G. Zhou. 2010. "Robust Portfolios: Contributions from Operations Research and Finance." *Annals of Operations Research* 176 (1): 191–220.

Feurer, M., and F. Hutter. "Hyperparameter Optimization." In *Automated Machine Learning*, edited by F. Hutter, L. Kotthoff, and J. Vanschoren, pp. 3–33. The Springer Series on Challenges in Machine Learning. Cham, Switzerland: Springer. 2019.

Fortin, F. A., F. M. De Rainville, M. A. Gardner, M. Parizeau, and C. Gagné. 2012. "DEAP: Evolutionary Algorithms Made Easy." *Journal of Machine Learning Research* 13: 2171–2175.

Grinold, R. C. 2006. "A Dynamic Model of Portfolio Management." *Journal of Investment Management* 4 (2): 5–22.

Grinold, R. C., and R. N. Kahn. *Active Portfolio Management: A Quantitative Approach for Providing Superior Returns and Controlling Risk*, 2nd ed. New York: McGraw–Hill. 2000.

Guidolin, M., and A. Timmermann. 2007. "Asset Allocation under Multivariate Regime Switching." *Journal of Economic Dynamics and Control* 31 (11): 3503–3544.

Gülpınar, N., and B. Rustem. 2007. "Worst-Case Robust Decisions for Multi-Period Mean–Variance Portfolio Optimization." *European Journal of Operational Research* 183 (3): 981–1000.

Hansen, N., and A. Ostermeier. 2001. "Completely Derandomized Self-Adaptation in Evolution Strategies." *Evolutionary Computation* 9 (2): 159–195.

Herzog, F., G. Dondi, and H. P. Geering. 2007. "Stochastic Model Predictive Control and Portfolio Optimization." *International Journal of Theoretical and Applied Finance* 10 (2): 203–233.

Ho, M., Z. Sun, and J. Xin. 2015. "Weighted Elastic Net Penalized Mean–Variance Portfolio Design and Computation." *SIAM Journal on Financial Mathematics* 6 (1): 1220–1244.

Hutter, F., H. H. Hoos, and K. Leyton-Brown. "Sequential Model-Based Optimization for General Algorithm Configuration." In *Learning and Intelligent Optimization*, vol. 6683, pp. 507–523. Lecture Notes in Computer Science Series. Berlin: Springer. 2011.

Igel, C., N. Hansen, and S. Roth. 2007. "Covariance Matrix Adaptation for Multi-Objective Optimization." *Evolutionary Computation* 15 (1): 1–28.

Jagannathan, R., and T. Ma. 2003. "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps." *The Journal of Finance* 58 (4): 1651–1683.

Kolm, P. N., and G. Ritter. 2019. "Dynamic Replication and Hedging: A Reinforcement Learning Approach." *The Journal of Financial Data Science* 1 (1): 159–171.

Kolm, P., R. Tütüncü, and F. Fabozzi. 2014. "60 Years of Portfolio Optimization: Practical Challenges and Current Trends." *European Journal of Operational Research* 234 (2): 356–371.

Ledoit, O., and M. Wolf. 2003. "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection." *Journal of Empirical Finance* 10 (5): 603–621.

——. 2004. "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis* 88 (2): 365–411.

Li, J. 2015. "Sparse and Stable Portfolio Selection with Parameter Uncertainty." *Journal of Business & Economic Statistics* 33 (3): 381–392.

Loshchilov, I., T. Glasmachers, and H. G. Beyer. 2019. "Large Scale Black-Box Optimization by Limited-Memory Matrix Adaptation." *IEEE Transactions on Evolutionary Computation* 23 (2): 353–358.

Markowitz, H. 1952. "Portfolio Selection." *The Journal of Finance* 7 (1): 77–91.

Meindl, P. J., and J. A. Primbs. 2008. "Dynamic Hedging of Single and Multi-Dimensional Options with Transaction Costs: A Generalized Utility Maximization Approach." *Quantitative Finance* 8 (3): 299–312.

Mulvey, J. M., and B. Shetty. 2004. "Financial Planning via Multi-Stage Stochastic Optimization." *Computers & Operations Research* 31 (1): 1–20.

Nelder, J. A., and R. Mead. 1965. "A Simplex Method for Function Minimization." *Computer Journal* 7 (4): 308–313.

Nystrup, P., S. Boyd, E. Lindström, and H. Madsen. 2019. "Multi-Period Portfolio Selection with Drawdown Control." *Annals of Operations Research* 282 (1–2): 245–271.

Nystrup, P., B. W. Hansen, H. O. Larsen, H. Madsen, and E. Lindström. 2017. "Dynamic Allocation or Diversification: A Regime-Based Approach to Multiple Assets." *The Journal of Portfolio Management* 44 (2): 62–73.

Nystrup, P., B. W. Hansen, H. Madsen, and E. Lindström. 2015. "Regime-Based Versus Static Asset Allocation: Letting the Data Speak." *The Journal of Portfolio Management* 42 (1): 103–109.

Nystrup, P., H. Madsen, and E. Lindström. 2018. "Dynamic Portfolio Optimization across Hidden Market Regimes." *Quantitative Finance* 18 (1): 83–95.

Pedregosa, F. 2016. "Hyperparameter Optimization with Approximate Gradient." *PMLR: Proceedings of the 33rd International Conference on Machine Learning* 48: 737–746.

Pınar, M. Ç. 2007. "Robust Scenario Optimization Based on Downside-Risk Measure for Multiperiod Portfolio Selection." *OR Spectrum* 29 (2): 295–309.

Powell, M. J. D. "A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation." In *Advances in Optimization and Numerical Analysis*, pp. 51–67. Dordrecht: Springer. 1994.

Snoek, J., H. Larochelle, and R. P. Adams. 2012. "Practical Bayesian optimization of machine learning algorithms." *Advances in Neural Information Processing Systems* 25: 2951–2959.

Stein, C. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 197–206. Berkeley: University of California Press. 1956.

Voß, T., N. Hansen, and C. Igel. "Improved Step Size Adaptation for the MO-CMA-ES." In *Proceedings of the 12th Annual Genetic and Evolutionary Computation Conference*, pp. 487–494. New York: ACM. 2010.

Wills, A. G., and T. B. Schön. "On the Construction of Probabilistic Newton-Type Algorithms." In *Proceedings of the 56th IEEE Annual Conference on Decision and Control*, pp. 6499–6504. New York: IEEE. 2017.

Zenios, S. A. *Practical Financial Optimization: Decision Making for Financial Engineers*. Malden, MA: Blackwell, 2007.

Zhang, R., N. Langrené, Y. Tian, Z. Zhu, F. Klebaner, and K. Hamza. 2018. "Dynamic Portfolio Optimization with Liquidity Cost and Market Impact: A Simulation-and-Regression Approach." *Quantitative Finance* 19 (3): 519–532.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Dynamic Replication and Hedging: *A Reinforcement Learning Approach*

Petter N. Kolm and Gordon Ritter

The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/1/159

ABSTRACT: *The authors of this article address the problem of how to optimally hedge an options book in a practical setting, where trading decisions are discrete and trading costs can be nonlinear and difficult to model. Based on reinforcement learning (RL), a well-established machine learning technique, the authors propose a model that is flexible, accurate, and very promising for real-world applications. A key strength of the RL approach is that it does not make any assumptions about the form of trading cost. RL learns the minimum variance hedge subject to whatever transaction cost function one provides. All that it needs is a good simulator, in which transaction costs and option prices are simulated accurately.*

Big Data in Portfolio Allocation: *A New Approach to Successful Portfolio Optimization*

Irene Aldridge

The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/1/45

ABSTRACT: *In the classic mean–variance portfolio theory as proposed by Harry Markowitz, the weights of the optimized portfolios are directly proportional to the inverse of the asset correlation matrix. However, most contemporary portfolio optimization research focuses on optimizing the correlation matrix itself, and not its inverse. In this article, the author demonstrates that this is a mistake. Specifically, from the Big Data perspective, she proves that the inverse of the correlation matrix is much more unstable and sensitive to random perturbations than is the correlation matrix itself. As such, optimization of the inverse of the correlation matrix adds more value to optimal portfolio selection than does optimization of the correlation matrix. The author further shows the empirical results of portfolio reallocation under different common portfolio composition scenarios. The technique outperforms traditional portfolio allocation techniques out of sample, delivering nearly 400% improvement over the equally weighted allocation over a 20-year investment period on the S&P 500 portfolio with monthly reallocation. In general, the author demonstrates that the correlation inverse optimization proposed in this article significantly outperforms the other core portfolio allocation strategies, such as equally weighted portfolios, vanilla mean–variance optimization, and techniques based on the spectral decomposition of the correlation matrix. The results presented in this article are novel in the data science space, extend far beyond financial data, and are applicable to any data correlation matrixes and their inverses, whether in advertising, healthcare, or genomics.*

A Network Approach to Analyzing Hedge Fund Connectivity

GUEORGUI S. KONSTANTINOV AND JOSEPH SIMONIAN

GUEORGUI S. KONSTANTINOV is a senior portfolio manager in fixed-income & currencies at LBBW Asset Management in Stuttgart, Germany.
georgi.konstantinoff@gmail.com

JOSEPH SIMONIAN is the founder and CIO of Autonomous Investment Technologies LLC in Newton, MA.
jsimonian@autoinvestec.com

KEY FINDINGS

- The hedge fund universe can be represented as a graph that depicts the relational structure and interaction among individual hedge funds.
- Clustering algorithms are used to classify and evaluate hedge fund styles as well as individual hedge funds. Hedge fund networks are found to have different dynamics across different stages of the business cycle.
- Various hedge funds are found to be connected based on factor exposure in ways that stand in contrast to their database categorizations; applying LASSO helps to identify factor exposure in the clusters and to predict network behavior.

ABSTRACT: *In this article, the authors investigate the hedge fund market as a network of interacting individual funds. The authors identify and analyze the most important hedge fund styles that could both affect the market and transmit systemwide shocks to other funds, individual asset classes, and beyond. The authors find that the most connected hedge fund database categories are global macro and equity long–short funds. A central result of the article is a classification of funds using clustering, in which seemingly different funds are shown to cluster based on their shared factor exposures. This finding demonstrates that investors should consider fund connectivity and their attendant importance scores rather than database classifications when measuring hedge fund risk across the business cycle. The authors also provide a forecasting framework that can be used to predict hedge fund network behavior and the impact of individual factors on the network.*

TOPICS: ***Analysis of individual factors/risk premia, factor-based models, style investing, performance measurement****

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

Hedge funds can be characterized as interacting agents in a network.[1] Although a large body of literature has examined hedge funds and hedge fund interconnectedness on an index level, only a handful of papers has considered interconnectedness of single hedge fund strategies and their underlying risk factors, which ultimately determine fund connectedness. As Provost and Fawcett (2013) argued, information about network structure is important and helpful in building models and predicting future values of risk factor exposure.[2]

[1] See Loistl and Landes (1989) and Landes and Loistl (1992) for a characterization of the financial system as an interacting agent-based network.

[2] This argument is related to the discussion of allocation decisions and the transmission of systemic and systematic risk as presented by Khandani and Lo (2011).

This article was motivated by the lack of research on single-fund connectivity and specific network metrics relevant for describing agent- or fund-based interconnectedness. To that end, we define a *network*, attributing a proper measure of risk to the nodes (hedge funds) and their relationships (mutual links). Specifically, at a high level, the hedge fund universe can be represented as a graph that characterizes the network of interacting individual hedge funds and information flow regarding returns and factor exposures. It is imperative to identify clusters in the hedge fund network because clusters might have factor exposure and network properties that differ from those of the network as a whole. Cluster analysis may also reveal how risk is transmitted not only between funds and groups but also beyond, to other asset classes. In this respect, cluster analysis helps provide more flexibility in database classification.

The primary aim of this article is to demonstrate a multistep approach to analyzing the global hedge fund network and its factor exposures in a connected, evolving financial market. We show the common factor exposures of the funds and how they are related to connectedness. Investigating interconnectedness and its structure shows how and the extent to which financial risk can be transferred—within categories, single funds, and across the whole hedge fund universe.

Detecting mutual relations is a matter of statistical inference. As a natural consequence, measuring connectivity requires specific analytical techniques, visualization, and metrics beyond variances and covariances. These metrics are crucial for understanding how risk propagates in a network. The approach described in this article offers fund managers, analysts, and investors a novel method to integrate models from social networks and factor analysis. The method described provides some clear risk management benefits—specifically, the ability to potentially avoid investing in strongly connected funds with highly correlated factor exposures that thus present significant concentration risk. A central finding in the article is that funds are clustered more heavily based on specific factor exposures rather than vendor-specific database categories. Another significant finding is that strong fund connectivity during lower absolute return cycles and economic slowdowns increases the transmission of risk and thus lowers diversification. Predictive models show that network links, fund connectivity, and risk propagation can be forecasted. Finally, we provide guidance on how to practically implement the article's findings within portfolio management.

HEDGE FUND DATA: TRANSPARENCY AND BIASES

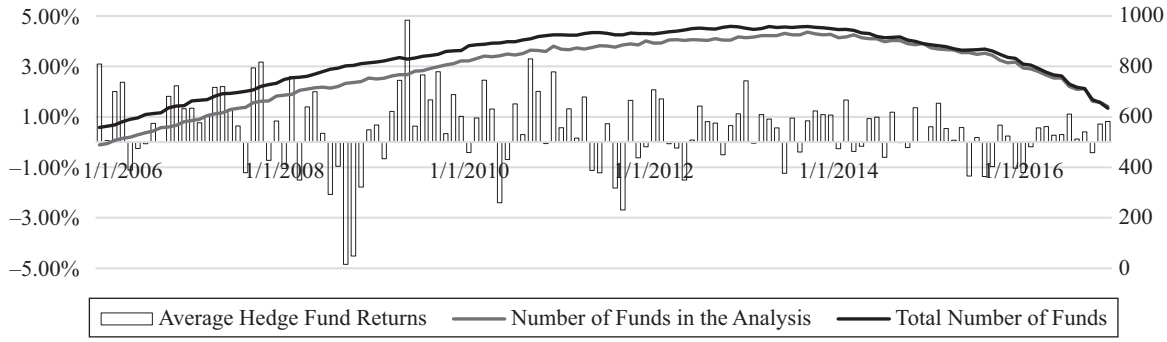
Malkiel and Saha (2005) and Ibbotson, Chen, and Zhu (2011) argued that hedge fund returns show serious survivorship bias and backfill bias, which affects performance and risk measures. Liang (2000) argued that different databases contain diverse numbers of dissolved funds. Moreover, the possibility of style drift is a common risk for the hedge fund industry, and it may not be feasible for a database provider to determine its extent. In this article, we draw on the MercerInsight hedge fund database. To begin, we briefly review how this database handles the foregoing challenges:

- *Backfill bias.* The database provides strong backfill regulation. With a backfilling rule in place to no longer include full historical performance for newly entered funds, track records are restricted to one quarter's back history in the database. In this article, we restrict the analysis only to funds with at least 12 months of reported returns.
- *Representativeness and fund attrition rate.* MercerInsight provides a track record of strategies rather than individual funds and only includes one performance track record per strategy for inclusion within the official universe. The sample shows that only 239 (or 40.7%) of the 587 reported on January 2006 survived the full sample period. These rates are consistent with the rates reported by Brown, Goetzmann, and Park (2001).
- *Survivorship bias.* This bias occurs when database vendors exclude the historical performance of defunct or closed funds from the database. Survivorship bias overestimates returns and underestimates risk. We consider the delisted funds and report survivorship bias of 134 bps.

The MercerInsight database reports unsmoothed, net-of-fees returns in the period from January 2006 to December 2016 (see Exhibit 1). Examination of data for duplicates, autocorrelation, and inconsistencies is essential. Adjusting for biases, a total of 1,473 hedge funds were available for analysis with a track record of more than 12 monthly returns, including 1,098

EXHIBIT 1

Average Hedge Fund Returns: January 2006–December 2016



Notes: The number of funds refers to funds that have at least 12 monthly observations—1,473 in total; the total number of funds in the database during the sample period was 1,689. MercerInsight customized hedge fund data.

dissolved funds. The total strategy assets were USD 1.27 trillion as of December 2016.[3]

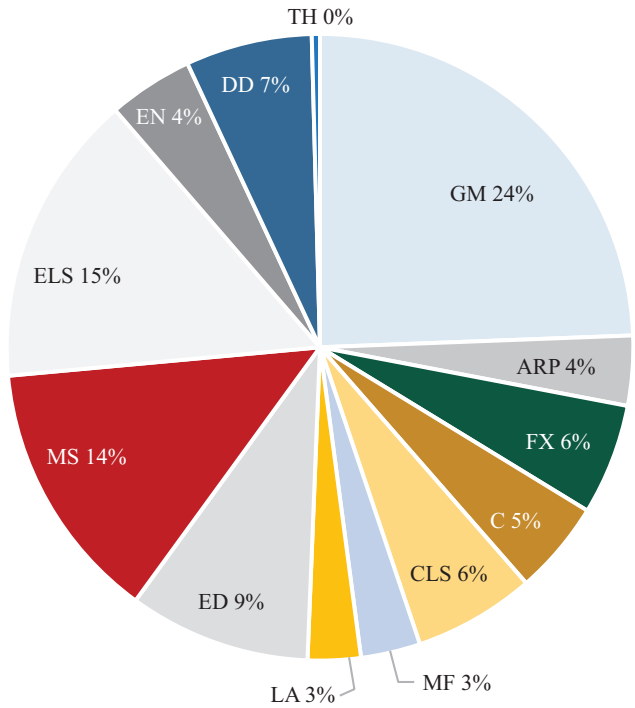
To frame the study, we use the fund categorization supplied by MercerInsight when reporting results that covers the following categories: Global Macro (GM), Alternative Risk Premia (ARP), Active Currency (FX), Commodities (C), Credit Long Short (CLS), Managed Futures (MF), Event-Driven (ED), Liquid Alternatives (LA), Multi-strategy (MS), Equity Long Short (ELS), Equity Market Neutral (EN), Distressed Debt (DD), and Tail Risk Hedge (TH). They exclude funds of hedge funds (FoHF) from the analysis because these differ in terms of fees, compensation, liquidity, leverage, and tactical allocation. Exhibit 2 shows the percentage of strategy assets in each corresponding category.

CAPTURING HEDGE FUNDS' INTERCONNECTEDNESS

Decomposing the drivers of return, risks, and common database biases for hedge funds has a long tradition, exemplified most notably in the works of Fung and Hsieh (1997, 2002); Schneeweis and Spurgin (1998); Liang (2001); Agarwal and Naik (2000); Capocci and Hübner (2004); Dor, Dynkin, and Gould (2006); Bollen and Whaley (2009); Blitz (2018); and Simonian and Wu (2019a). In addition to the latter research, there has also

EXHIBIT 2

Strategy Assets According to the Database Classification



Notes: MercerInsight customized hedge fund data. AUM reported in billions (USD) as of December 2016.

been a recognition that network analysis can be used as an effective tool to analyze hedge fund connectedness. López de Prado (2016) for example, has argued that graph theory successfully explains complex relationships.

[3] The total hedge fund industry assets under management (AUM) are hard to estimate, and the number of hedge funds exceeds 8,377 according to Kazemi, Black, and Chambers (2016). In fact, 1,419 of the funds considered in the analysis reported their strategy assets.

Getmansky, Lee, and Lo (2015) investigated hedge funds and money market funds and also stressed the usefulness of graph theory in investigating relations. Baitinger and Papenbrock (2017) investigated extensive analysis on interconnectedness in the active investment management context.[4] Simonian and Wu (2019b) showed that graph theoretic approaches to clustering can be used to analyze macro regimes.

In this line of thought, Billio et al. (2012) stressed that funds within the same category might transmit systemic shocks as a result of style similarity. To show how hedge funds move with common trends, we run a panel cointegration test. Cointegrated series are linearly related in a stationary process. The common average (stationary) value may well indicate a common factor exposure. We are able to reject the null hypothesis of no cointegration (augmented-Dickey–Fuller test statistic −46.111; $p = 0.000$). Thus, the test clearly suggests that the fund returns are cointegrated and converge to a linear stationary term.

For our factor analysis, we use a 10-factor model composed of equity and fixed-income components. For the equity component of our model, we the Fama–French (1992, 1993, 2015) five-factor model factors—the equity premium (Mkt-Rf), value (HML), size (SMB), profitability (RMW), investment (CMA)—momentum (Carhart 1997; Moskowitz, Ooi, and Pedersen 2012), and an illiquidity factor that we define as the one-month lagged equity premium ME($t - 1$).[5] For the excess returns on long–short equity we use betting-against-beta factors (BAB), which are portfolios that are long low-beta securities and short high-beta securities (Frazzini and Pedersen 2014).[6] For our fixed-income factor, we use the yield curve spread, measured as the month-end to month-end difference of the BBB Corporate Yield and the 10Y US Treasury Bond Yield (YCSpread). We formally express our model in the familiar manner:

$$R_{i,t} = \alpha_i + \sum_{i=1} \beta_i \; F_{j,t} + \varepsilon_t \tag{1}$$

where $R_{i,t}$ is the total return generated by the fund, α is the measure of manager skill, β is a coefficient that measures the sensitivity of a fund's return to the factor, F is the beta factor that requires a systematic risk premium in the market, and ε is the random error term. We report the descriptive statistics for the risk factors in Exhibit A1 in Appendix A.[7]

MODEL SPECIFICATION AND HEDGE FUND CLASSIFICATION

The lack of transparency in the hedge fund industry and the difficulties of gathering data—less public information regarding portfolio holdings, AUM, and weights and the need for estimating relevant metrics that efficiently capture fund relationships—necessitate alternative methods for measuring connectedness. Konstantinov and Rebmann (2019) and Konstantinov and Rusev (2020) proposed a method using fund returns. Funds may have different than expected systematic exposures. Following this line of thought, return connectedness can be captured using an approximating classification model, if the funds are converging toward similar style exposure, as shown by what is known as *view commonality*. Thus, view commonality and cointegration together enable a comparison among alternative funds and portfolios that have common factor exposures.

The factor exposures of a fund are the weighted average of its assets' factor exposures. If two funds are connected, their link is determined by their factor exposure. What is known as an *adjacency matrix* captures the pairwise links (Das 2016). As in the work of Sandhu, Georgiou, and Tannenbaum (2016), connectivity exists under the assumption of fund interactions[8]

[4] Earlier, Bussiere, Hoerova, and Klaus (2014) discussed methods of measuring connectedness in the hedge fund industry. In fact, having the same phenomenon in mind, Khandani and Lo (2011) referred to graph theory and used the term *connectedness*. Furthermore, Khandani and Lo (2011) and Adrian (2007) referred to co-movement and style similarity, and Pojarliev and Levich (2011) used the term *crowdedness*.

[5] An earlier treatment of risk factors using machine learning can be found in Simonian et al. (2019); the authors applied the random forest algorithm to the Fama–French–Carhart model.

[6] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, https://www.aqr.com/Insights/Datasets/Value-and-Momentum-Everywhere-Factors-Monthly, and https://www.aqr.com/Insights/Datasets/Betting-Against-Beta-Equity-Factors-Monthly.

[7] Unreported rolling regressions using Equation 1 show that the common factor exposure (view commonality) is normally distributed over the sample period with an average value of 31.09% (median 30.17%) and a standard deviation of 9.63%. The model explains roughly 61% of the variance of hedge fund returns in the sample period.

[8] This was also stressed most recently by Loistl and Konstantinov (2020).

Fund interaction is determined on the risk factor exposure. Thus, the network is risk factor based.

A linear model is suitable to capture hedge fund interconnectedness:

$$R_{i,t} = \alpha_i + \beta_i \; r_{j,t} + \varepsilon_t \tag{2}$$

In a classification algorithm, it is not the magnitudes of the coefficient β_i that are relevant, but solely their significance. Here, $R_{i,t}$ is the return of fund i, regressed against the returns of fund j at time $t(r_{j,t})$ and against the squared returns $r^2_{j,t}$ of fund j. We consider the significant coefficients in Equation 2 to fill the elements of an (adjacency) matrix mapping of the pairwise links between the funds continuously through the sample period, with a time step of one monthly observation, using a rolling window of 12 observations. The t-statistic is 2.733, which is adjusted for false discoveries.

INTEGRATING FACTOR ANALYISIS AND NETWORK MODELS

Now that we have set up our factor model, we proceed to investigate how it can be integrated with network models. Constructing networks in a dynamic manner allows information feedback to be incorporated into the analysis of the hedge fund universe, as opposed to merely showing average relationships between funds.[9] Furthermore, the metrics capturing connectedness are otherwise not observable using standard statistics, such as variance and covariances.

As mentioned previously, detecting clusters is central for networks. Modularity is one such measure, applied to network community (cluster) detection on a node level. Tests on single hedge fund correlation networks show that the modularity is lower than that for a directed network. Hedge funds represent a directed network. Specifically, hedge fund managers have sources of information regarding their factor exposures that may be unbiased by the factor views of other fund managers. Thus, a hedge fund might be linked to another fund, but that does not mean this second fund is linked back to the first one. In other words, the relationship may be asymmetric and would thus not be captured by a measure such as correlation.[10,11]

Defining the structure of a directed network requires specific metrics. The most basic are the number of *nodes n* (funds) and *links* (connections) *m*. The number of nodes (funds) available represents the size of the network. The links are identified by pairs (i,j) and measure the relationship (connection) between nodes i and j (funds). An $n \times n$ *adjacency matrix*, denoted by A, is estimated using the following rule:

$$a_{ij,t} = \begin{cases} 1 \; if \;\; Fund \;\; i \;\; is \;\; connected \;\; to \;\; Fund \;\; j \\ 0 \;\; Otherwise \end{cases} \tag{3}$$

where a_{ij} are the components of the adjacency matrix and take a binary value. The value is 1 only if fund i is linked to fund j. If fund i does not have a significant influence on fund j, as measured by the significance of the coefficients of the regression model, the value in the matrix is set at zero.

The *in-degrees* and *out-degrees* are representations of the connections of a fund (node) and referred to as a measure of connectedness, as described by Newman (2010). In this case, this number shows the average connectedness of a fund, as measured by the adjacency matrix.

$$k^{in-degree}_{i,t} = \sum_i a_{ij,t} \text{ and } k^{out-degree}_{j,t} = \sum_j a_{ji,t} \tag{4}$$

The *density* of a network α expresses the ratio between the numbers of current links and the number of all possible links between the funds in the network at time t.

$$\alpha_t = \frac{m_t}{n_t(n_t - 1)} \tag{5}$$

Here, m represents the number of links in the network of managers, and n is the number of funds available. The density ranges from 0 (no fund ties are present) to 1 (completed network).

[9] Baitinger and Maier (2019) previously built correlation networks for hedge fund indexes.

[10] Regarding correlation measures, we further note an observation by Raffinot (2018) that correlations make no qualitative differentiation between assets, which is disadvantageous for hedge fund analysis given that individual hedge funds may show similarities to individual asset classes.

[11] We further note that Puliga, Caldarelli, and Battiston (2014) showed that correlation networks should be used with caution when analyzing systemic risk.

Reciprocity shows the relation between the pairs of funds in a network. A fund pair (i,j) is called reciprocal if ties exist between both funds in both directions. The values for reciprocity at time t range between 0 and 1 (with 1 for a fully reciprocal network):

$$\varrho_t = \frac{\sum_{ij} a_{ij,t} \; a_{ji,t}}{m_t}. \tag{6}$$

Centrality Measure

Suggested by Bonacich and Lloyd (2001), the alpha centrality measure has the following form, which assigns importance scores to the funds:

$$x = alpha \; A^T x + e \tag{7}$$

where x is the vector of centrality scores, vector e reflects exogenous effects of the external factors, and *alpha* reflects the relative importance of endogenous versus exogenous factors. Matrix I is the identity matrix, and A^T is the transposed adjacency matrix from Equation 3.

Within Exhibit 3, Panel A, we show the average degree, which indicates the mutual interconnectedness on the node level. It has a value of 94, which indicates that a single fund is connected to 94 other funds on average. Reciprocity, shown in Exhibit 3, Panel B, refers to the connectivity as mutual links between the funds. An observant reader can easily identify the positive relationship between returns and average degree. Of $n(n-1)$ possible combinations as captured by the density, on average, 12.1% are statistically significant. Within Exhibit 3, Panel C, we show the density of the network and within Exhibit 3, Panel D, the descriptive statistics. On the contrary, mutual links (reciprocity) between funds rise as returns decrease. The latter is a sign of the possible transmission of risks in the network. We further discuss these points in the following.[12]

CENTRALITY AS A MEASURE OF IMPORTANCE

We investigate the most connected funds in the full sample and in the two subsamples using centrality scores and degree scores. Kinlaw, Kritzman, and Turkington (2012) highlighted the centrality measure in determining the vulnerability of portfolio components to systemic risks and shocks. Generally, by measuring the centrality scores of a network, there will always be a fund with a higher score than the rest. Degree centrality identifies an important fund, according to the number of links it has. However, the degree centrality we show in Exhibit 4, Panel B, is simply the average score of the most central fund and measures the immediate effects and influence of factor exposures.

The alpha centrality measure developed by Bonacich and Lloyd (2001) assigns a score to every fund that is proportional to the sum of the scores of its neighbor funds. Furthermore, an alpha centrality score is appropriate for mapping asymmetric relationships in directed networks. Hedge fund managers have sources of information regarding their factor exposure that are independent of other fund managers, meaning that hedge funds have asymmetric connections. Specifically, the hedge fund network is consistent with a system in which each fund is linked to its neighbors simultaneously.[13] Thus, alpha centrality allows every fund manager to receive an unbiased importance score in the network (Exhibit 4, Panel A). Economically, a fund's importance score increases if a fund has connections to other funds that are themselves connected and thus important in a competitive environment. The centrality score provides information such as that a fund's long-term equilibrium factor risk is a function of the factor exposure of the funds to which it is related.

However, importance scores change dynamically as the hedge fund market evolves. Thus, estimation of centrality scores, monitoring, and factor exposure estimations would provide detailed information to investors and risk managers regarding single node risks for the whole network. The positive network correlation coefficient of −0.309 for the funds in the full sample period means that strongly connected funds are more likely to be connected to weakly linked funds (and follow similar factor exposure) than to other strongly connected funds. This probability remained stable with time, with a correlation coefficient of −0.2894 in the first sample period and network correlation coefficients of −0.3253 in the second subsample.

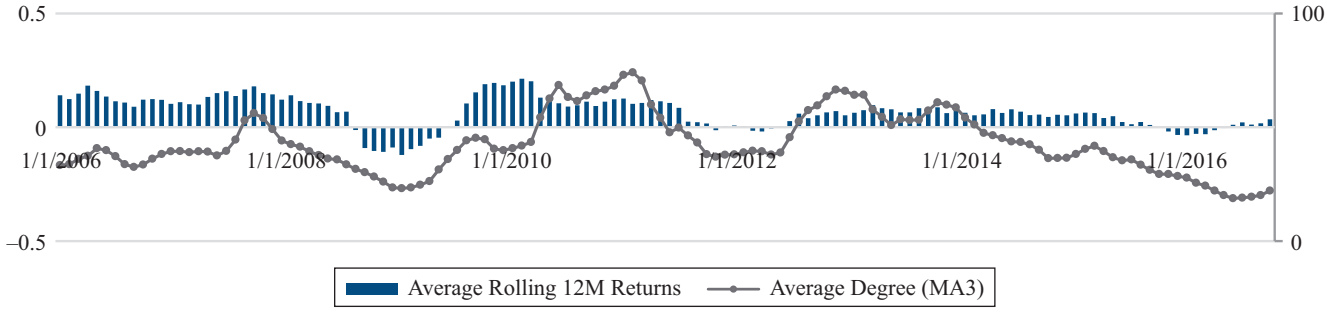
[12] We note that our reported degree and reciprocity values are similar to the ones reported by Bech and Atalay (2011) for the federal funds market.

[13] Borgatti (2005) referred to the specific centrality measures applicable to the different types of networks.

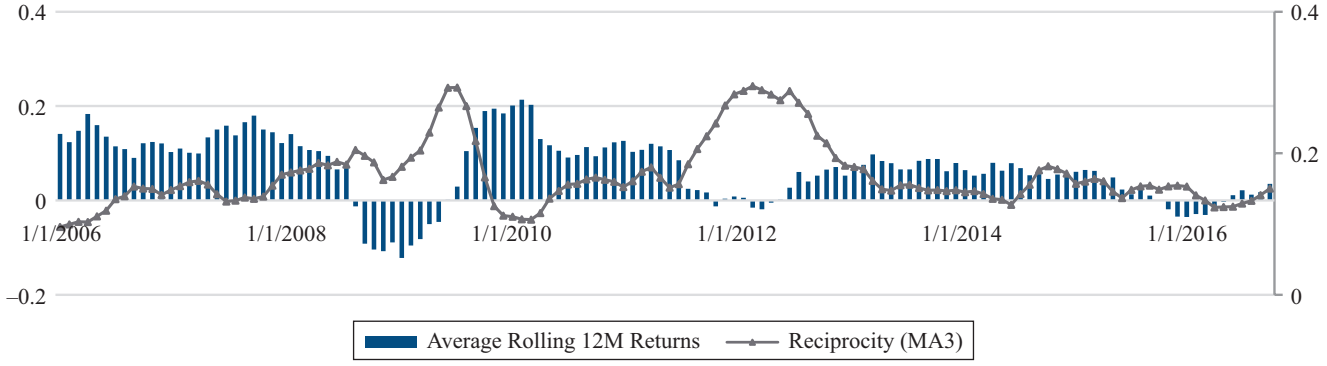
EXHIBIT 3

Network Metrics for Interconnectedness vs. Rolling Hedge Fund Returns: January 2006–December 2016

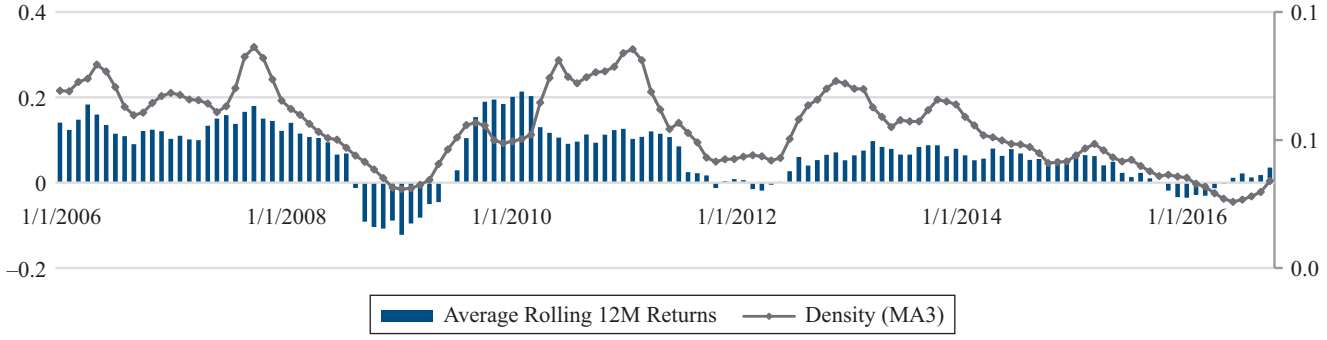
Panel A: Average Degree



Panel B: Reciprocity



Panel C: Density



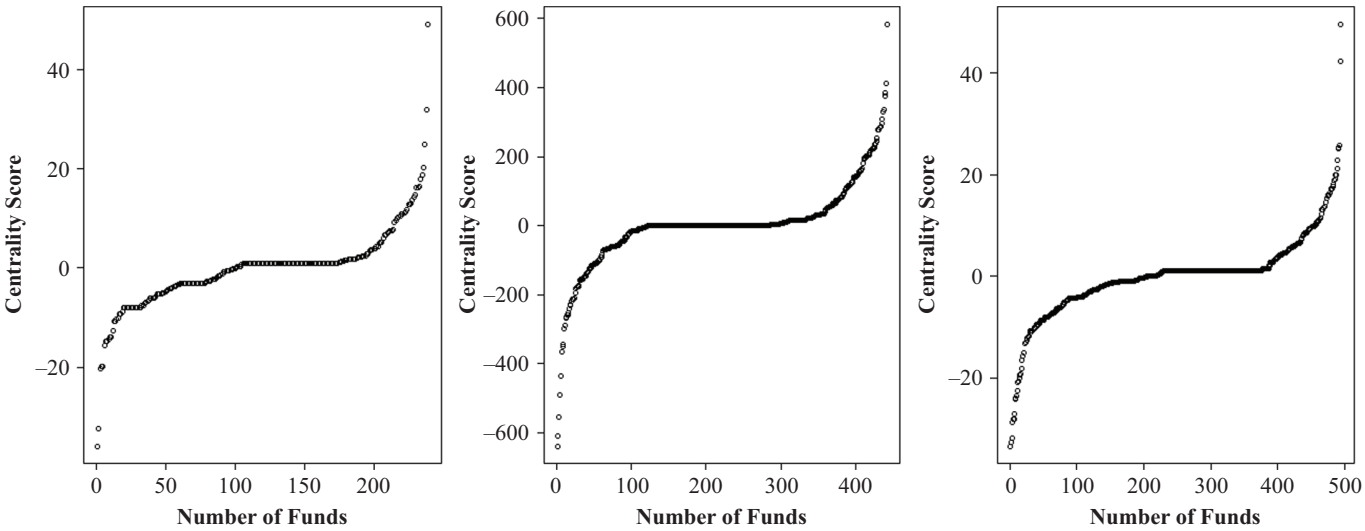
Panel D: Descriptive Statistics for the Network Metrics

	Average Degree	Reciprocity	Density
Mean	94.1259	0.1442	0.1210
Median	87.6164	0.1328	0.1159
Maximum	230.1128	0.2848	0.2823
Minimum	22.6292	0.0706	0.0315
Std. Dev.	50.1636	0.0541	0.0649
Skewness	0.5200	0.6390	0.4893
Kurtosis	2.3923	2.4800	2.2859

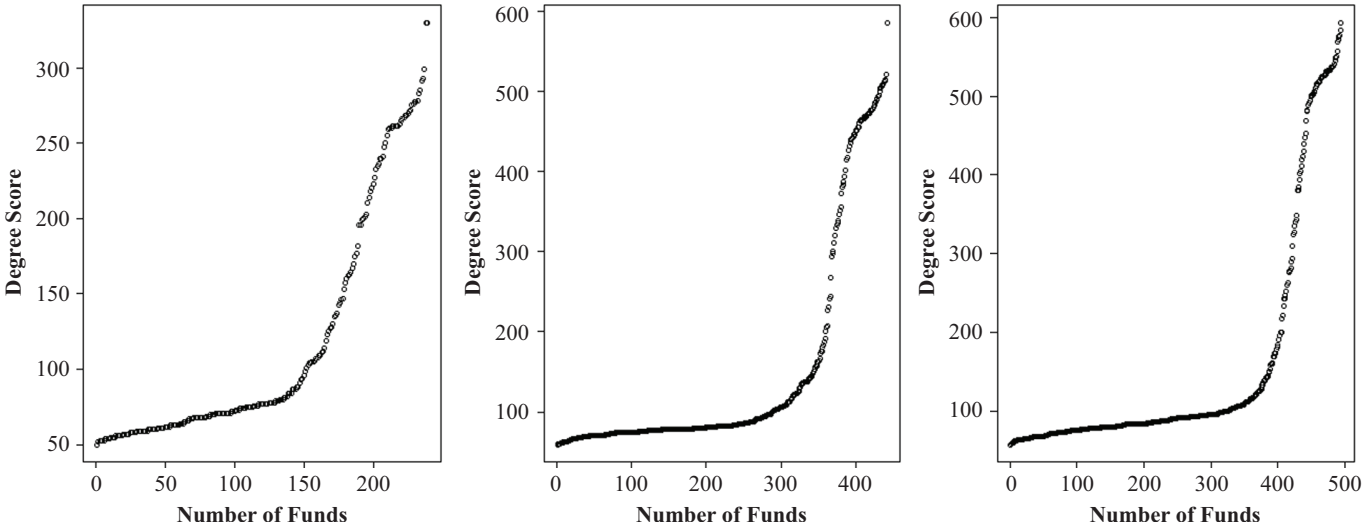
Notes: The first point is estimated using the first 12 monthly returns and corresponds to January 2006, whereas the last estimation is made on December 2016, with a rolling window of 12 observations and a time step of one month. The sample contains 132 rolling windows.

EXHIBIT 4
Centrality Scores

Panel A: Alpha Centrality



Panel B: Degree Centrality



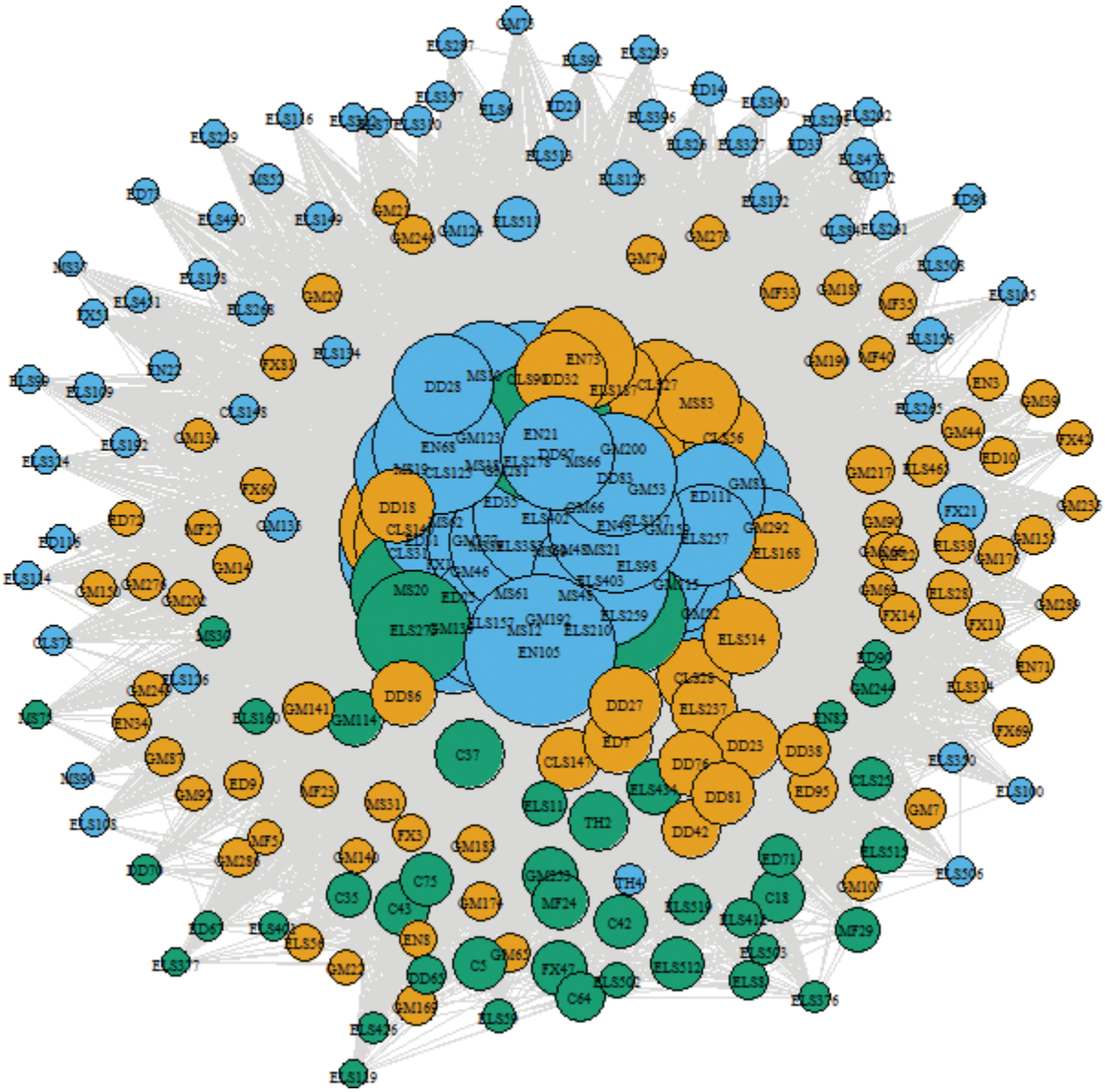
Notes: Alpha centrality is an absolute score. The degree centrality shows the number of funds to which a node (fund) is connected; MercerInsight customized hedge fund data.

CLUSTERS, INTERCONNECTEDNESS, AND RISK FACTORS

Clustering algorithms are widely used to classify data into distinct groupings. They are examples of unsupervised learning and thus do not require the labeling of input data before they is analyzed. Clusters enable risk factor analysis, based on funds' connectedness. Clusters are generally understood to be groups of funds that are more densely interconnected with each other based on underlying properties than with the rest of the funds in the network. Specifically, a GM, an FX, and an ELS fund might be in the same cluster, regardless of the database category to which they belong (see Exhibit 5). We apply the appropriate algorithm for large networks. The larger the dots, the higher the importance of the fund, according to its interconnectedness.

Exhibit 5

Interconnectedness of Hedge Funds with Full Track Record: January 2006–December 2016



Note: Adjacency matrix of the 239 funds with full track records in the sample period, the degree centrality scores as node weights, and the spinglass algorithm for community structure in large networks uses the framework from Yang, Algesheimer, and Tessone (2016).

The algorithm groups together funds that are strongly linked and separates those that are not. We find that clustering indicates a very different interconnectedness among the funds compared to their database categorization. This allows an understanding of the relationship between funds with similar properties that may not be possible from direct empirical tests or on a prior database categorization. Our results thus support the most recent findings of Baitinger and Maier (2019) that hedge fund strategies might misbehave with regard to their categorization.

The densest clusters comprise different types of funds; most importantly, there are different factor premiums, as indicated by the Fama–Macbeth (1973)

EXHIBIT 6

Fama–MacBeth Regressions and Cluster Analysis

	Intercept	BAB	CMA	EM.US	HML	ME(t – 1)	Mkt.RF	MOM	RMW	SMB	YCSpread
Panel A: January 2006–December 2016											
Cluster 1	0.0043	0.0009	0.0006	0.0035	0.0012	0.0091	0.0040	0.0056	–0.0015	0.0015	–0.0289
p-value	*0.0000*	*0.7837*	*0.6211*	*0.0552*	*0.2457*	*0.0097*	*0.0026*	*0.0161*	*0.1506*	*0.2093*	*0.1238*
Cluster 2	0.0041	–0.0033	–0.0028	0.0072	–0.0013	0.0094	–0.0026	0.0037	–0.0009	0.0042	–0.0405
p-value	*0.0000*	*0.2728*	*0.0841*	*0.0533*	*0.5442*	*0.0147*	*0.3371*	*0.0018*	*0.2182*	*0.1307*	*0.0276*
Cluster 3	0.0048	0.0021	–0.0022	–0.0046	–0.0001	0.0060	0.0027	0.0134	–0.0035	0.0002	–0.0060
p-value	*0.0001*	*0.8498*	*0.3580*	*0.3870*	*0.9745*	*0.4649*	*0.2769*	*0.0713*	*0.1507*	*0.9330*	*0.8882*
Panel B: April 2011–December 2016											
Cluster 1	0.0025	0.0198	0.0008	–0.0124	0.0024	–0.0019	0.0030	0.0107	0.0019	–0.0072	–0.0617
p-value	*0.0015*	*0.0004*	*0.5205*	*0.0011*	*0.2899*	*0.8444*	*0.1825*	*0.0003*	*0.4280*	*0.0749*	*0.0326*
Cluster 2	0.0047	0.0107	–0.0072	0.0048	0.0002	0.0659	–0.0136	0.0016	0.0026	0.0126	–0.0636
p-value	*0.0133*	*0.1693*	*0.0506*	*0.5866*	*0.9656*	*0.0602*	*0.0471*	*0.4494*	*0.2720*	*0.0541*	*0.0385*
Cluster 3	0.0037	–0.0006	–0.0008	0.0006	–0.0004	–0.0049	0.0035	–0.0017	–0.0006	–0.0001	–0.0047
p-value	*0.0000*	*0.8495*	*0.1922*	*0.6652*	*0.6611*	*0.1526*	*0.0004*	*0.2376*	*0.2487*	*0.9322*	*0.5325*
Panel C: January 2006–March 2011											
Cluster 1	0.0065	0.0040	–0.0010	0.0064	0.0005	0.0048	0.0042	0.0021	–0.0003	0.0016	–0.0119
p-value	*0.0000*	*0.0028*	*0.1856*	*0.0000*	*0.5806*	*0.0808*	*0.0026*	*0.0345*	*0.6268*	*0.0926*	*0.3785*
Cluster 2	0.0042	–0.0008	0.0005	0.0060	0.0015	0.0117	0.0020	–0.0005	–0.0005	0.0029	–0.0445
p-value	*0.0000*	*0.7653*	*0.6459*	*0.0239*	*0.2915*	*0.0023*	*0.1396*	*0.8323*	*0.6285*	*0.0690*	*0.0681*
Cluster 3	0.0065	0.0009	–0.0001	0.0057	0.0036	0.0090	0.0006	0.0075	–0.0026	0.0017	–0.0346
p-value	*0.0000*	*0.8602*	*0.9758*	*0.3801*	*0.0289*	*0.2960*	*0.8288*	*0.0069*	*0.1144*	*0.5770*	*0.3500*

Note: p-*values in italics; cluster affiliation according to the spinglass algorithm for directed networks.*

regressions in Exhibit 6. For example, the liquidity ME(t − 1), Mkt.RF, and MOM factor premiums are significant for Cluster 1 (Exhibit 6, Panel A). In the sample period between January 2006 and March 2011, Cluster 2 shows distinctive ME(t − 1), and Cluster 1 has a significant BAB premium. Consequently, hedge funds comprising clusters have significant factor exposures, weakly overlapping with other clusters (e.g., MOM for Cluster 1 and Cluster 3 in Exhibit 6, Panel C) because strategy definitions and exposure depend on the different database providers. Highly connected funds with a significant overlap of factor exposures are grouped into clusters. To generate predictions of the cluster membership, we use least absolute shrinkage and selection operator (LASSO).

PRACTICAL APPLICATIONS

In their seminal analysis, Feng, Giglio, and Xiu (2019) applied a quantitative process using a double LASSO to determine the significance of the factors in an empirical analysis. James et al. (2013) provided extensive analysis of machine learning techniques and LASSO implementation.

In Exhibit 7, Panel A, we see that whereas Cluster 1 might be classified as CMA and SMB factor driven, Clusters 2 and 3 show differences in MOM exposure and a different magnitude of CMA, EM.US, and Mkt.RF exposure. In the period from April 2011 to December 2016, Cluster 1 is EM.US and BAB driven. Clusters 2 and 3 are liquidity and Mkt.RF, HML, and MOM driven, respectively (Exhibit 7, Panel B). The distinctions between the clusters in Panel C in Exhibit 7 are easily identified.

We also investigate the dynamic relationship among network metrics, view commonality, and the business cycle, as measured by the change in first and second derivatives of the Morgan Stanley Global Co-Incidence Indicator.[14] We show our results in Exhibit 8, Panel A. As the exhibit shows, we find that in general

[14] This indicator is also used by Konstantinov and Rebmann (forthcoming).

EXHIBIT 7
LASSO Regressions and Cluster Analysis

Intercept	Mkt.RF	SMB	HML	RMW	CMA	MOM	YCSpread	ME(t – 1)	EM.US	BAB	Intercept
Panel A: January 2006–December 2016											
Cluster 1	0.0050	0.2659	0.0346	–	–	–0.2209	–	–0.0089	0.0292	–	–
Cluster 2	0.0033	0.1770	–	–	–	–0.0379	0.1869	–	0.0367	0.0333	–
Cluster 3	0.0042	0.2214	–	–	–0.0817	–0.1479	–0.0250	–0.0055	0.0424	0.1072	–
Panel B: April 2011–December 2016											
Cluster 1	0.0024	0.2011	–	–	–	–	–	–	–	0.0444	–0.0722
Cluster 2	–0.0004	0.0333	–	–	–0.0217	–	0.0874	–	0.0439	–	–
Cluster 3	0.0025	0.1938	0.0242	–0.0769	–0.2096	–	0.1057	–	0.0058	–	–
Panel C: January 2006–March 2011											
Cluster 1	0.0093	0.1415	–0.0474	–0.1510	–0.4137	–0.2113	0.0569	–0.0113	0.0856	0.0696	0.0826
Cluster 2	0.0070	0.2152	0.0454	–	–	–0.2678	–	–0.0111	0.0473	0.0298	0.0691
Cluster 3	0.0095	0.0888	–	–	–0.2066	–0.2921	–	–0.0093	0.0764	0.0347	0.0564

Note: Cluster affiliation according to the spinglass algorithm for directed network and cross-validations.

the average degree decreases during contraction, when the average hedge fund returns are negative. Conversely, we find that it increases strongly during recovery and expansion as returns recover. Not surprisingly, during expansion the connectivity (reciprocity) rises as returns peak. Nonetheless, during slowdown, as returns are still high, the single node connectedness peaks and the mutual links are still elevated. A possible explanation is that, during an economic slowdown, connectedness increases at a decreasing rate. However, the transition from slowdown to expansion and vice versa poses the greatest risk. In fact, the absolute value of every connectedness metric is highest during slowdown and lowest during recovery (see Exhibit 8, Panel B).

What do these results tell us about hedge fund investability? Our first observation is that the funds with the lowest importance score can make a valuable contribution in an FoHF portfolio. Constructing an FoHF with the methodology described could proceed as follows:

1. Detect common risk factor exposure.
2. Identify clusters.
3. Run predictive regressions, regularization, and determine feature importance.
4. Determine the business cycle dependence of the network and predict links.
5. Monitor the changing dynamics of factor exposure, risk, and connectedness.
6. Simulate and predict returns, and add (remove) nodes as the hedge fund system evolves and new (old) funds emerge (disappear).

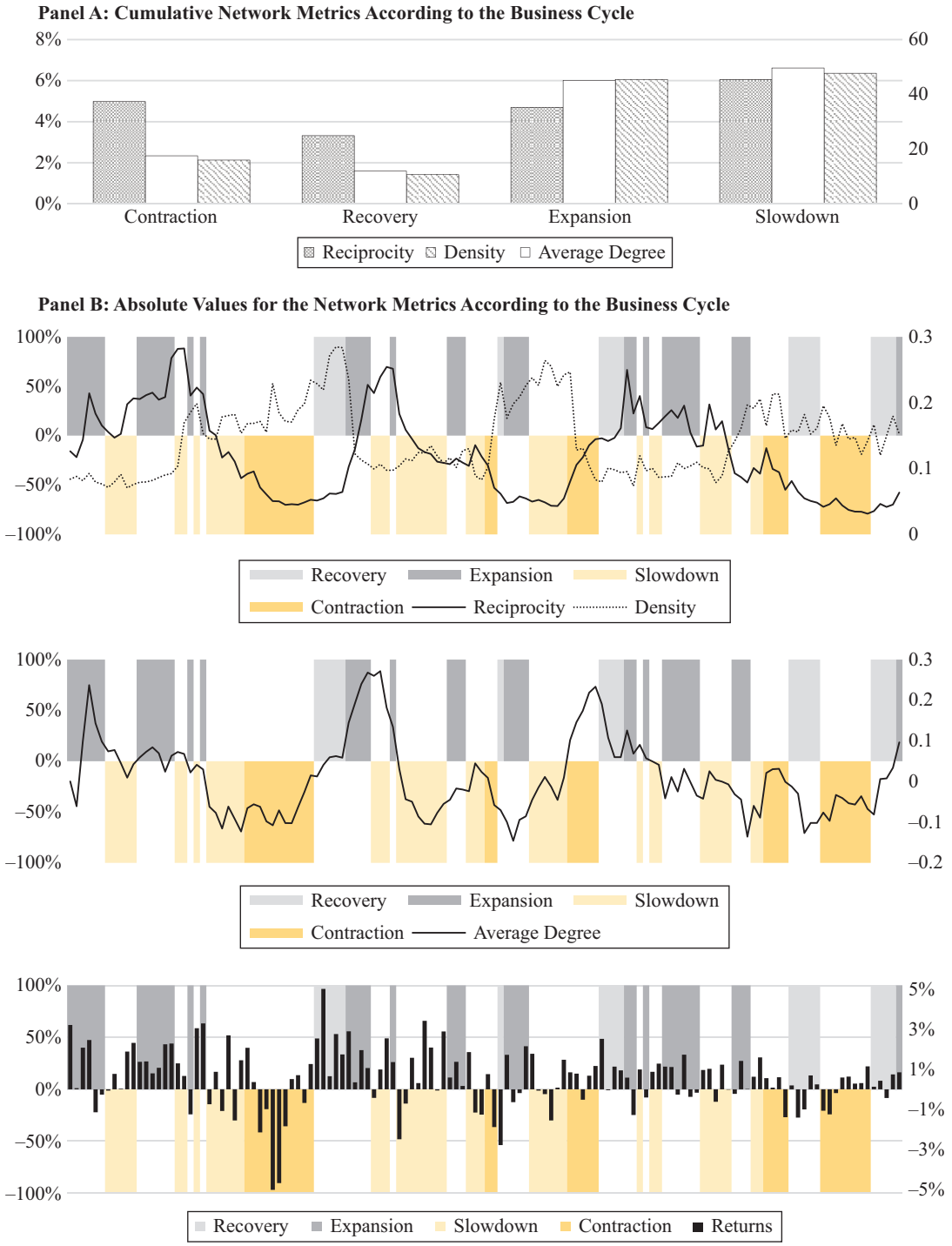
PREDICTING RISK IN HEDGE FUND NETWORKS

Factor exposures reflect risk and portfolio management as well as the important role of hedge funds in financial markets. Therefore, decomposing hedge fund network metrics into components allows for a direct estimation of common systematic risk factors on a network level. To show that the hedge fund network is factor driven, we apply LASSO cross-validation to the main network properties. We show the output in Exhibit 9.

The results clearly indicate that the network links are strongly driven by Mkt.RF (0.246), SMB (0.068), MOM (0.892), and the liquidity factor ME(t − 1) (0.574). Similar exposure to the four factors determines the changes in the node-level connectedness (average degree). The overall network density is to a large extent driven by the common factors, too. Importantly, the liquidity factor ME(t − 1) is significant for the network. Interestingly, there is a negative relationship between the changes in the network metrics and the CMA factor.

EXHIBIT 8

Dynamics of Business Cycle, Connectedness, and Hedge Fund Returns



Notes: We compute the phases of the business cycle based on the level of change and the second derivative of the Morgan Stanley Global Coincident Indicator, and the levels of the average degree, density, reciprocity, and average fund returns. The level of 3.32% is used to set the threshold between slowdown/contraction and recovery/expansion and is in line with the recent average. Average degree value is plotted on the right-hand side in Panel A.

EXHIBIT 9

Predictive Regressions for the Network: January 2006–December 2016

	(Intercept)	Mkt.RF	SMB	HML	RMW	CMA	MOM	YCSpread	ME(t – 1)	EM.US	BAB
Links	0.006	0.246	0.068	–	–	–0.256	0.892	–	0.588	0.105	–
Density	–0.001	0.203	0.013	–	–	–0.755	0.930	–	0.575	–	–
Reciprocity	0.047	–	–	–	–	–	–	–	–	–	–
Av. Degree	0.002	0.234	0.063	–	–	–0.566	0.906	–	0.574	0.006	–

Das, Kim, and Ostrov (2019) investigated graph properties, simulating random networks while increasing the probability of equal links. Simulation shows the development of the network, holding the number of links constant, or simulating the links for a constant number of nodes. The results, based on random graph generation, confirm previous findings; there is less than a linear relationship between the percentage increase in probability of two funds being linked (reciprocity), holding the number of funds constant. Increasing the number of funds with a constant number of mutual edges constitutes an increase in systemic risk. Funds managed by the same manager—depending on their size—might have a significant impact on the total network and bear systemic risk. This can be easily verified by simulating a network while increasing the number of both nodes and links.[15] The rise in the node-level connectivity (average degree) increases with an increasing rate, whereas the percentage rise in links in the network increases with a decreasing rate. For a simple example, consider two funds listed as different share classes that are actually managed by the same manager. Having different funds managed by the same manager, with similar factor exposures, does not reduce overall network risk. On the contrary, risk and factor connectedness increases as the fund connectedness increases.

Kundro and Feffer (2004) found that roughly 58% of hedge fund failure is attributable to operational risk and roughly 38% to market risk factors. Operational risk refers to economic outcomes resulting from inefficient implementation of investment decisions. Cluster factor analysis, such as the approach presented in this article, can validate the need to enhance the analysis to investigate both financial and operational risks. Strongly connecting underperforming funds with excess volatility, inefficient factor bets, and high attrition probability pose serious idiosyncratic and systemic risks. Due diligence remains an important dimension that the current analysis cannot replace.

As an additional robustness test, we run factor regressions using the Fama–French–Carhart six-factor model. We report average view commonality of 53.61% with a standard deviation of 10.56% (average R^2 of 0.72). Importantly, a time-series cross-validation, adjusting for purging and embargo, is used to validate results.[16] In Exhibit 10, we see that this analysis produces a correlation network of five distinctive clusters.

CONCLUSION

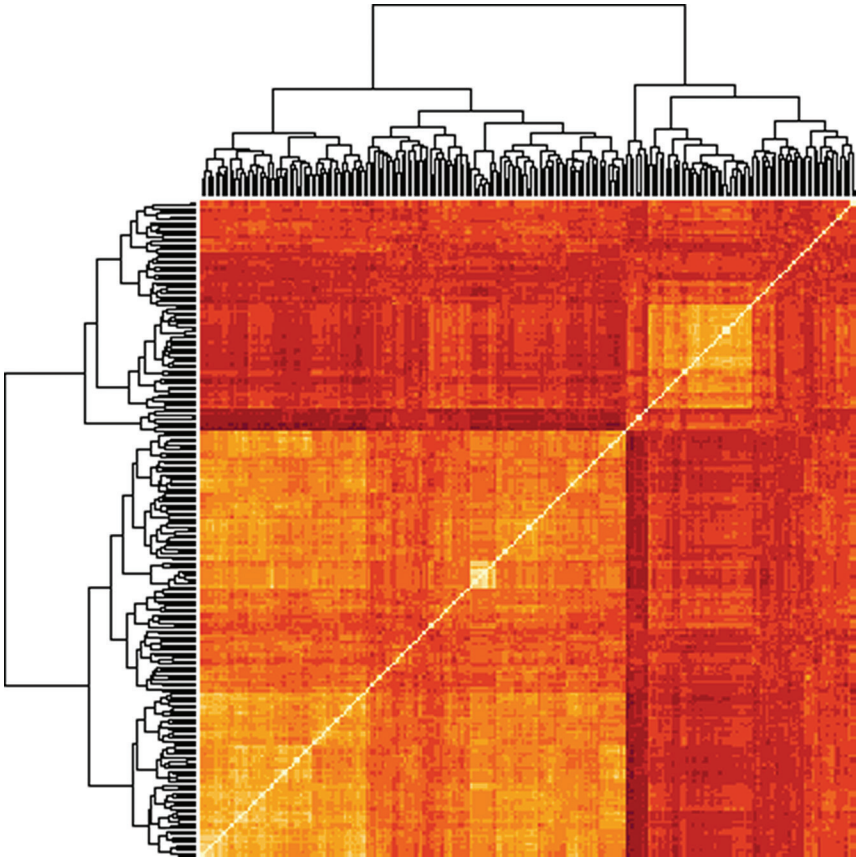
In this article, the structure and connectedness of hedge funds is analyzed using graph theoretic methods. We perform factor analysis and detect the common exposure of hedge funds in the sample period of January 2006–December 2016. The central finding of the study is that the database classification is loose; single funds are much more interconnected within clusters, driven by common factor exposures with distinctive factor premiums in the market, momentum, and size factors. There are important practical implications for both portfolio managers and risk managers interested in alternative assets. Given the crowded behavior in assets and factors and the sensitivity of hedge funds, the most connected funds, which are linked to other strongly connected funds, are the bridges through which the web can be affected most in the event of market turmoil caused by both exogenous and endogenous shocks. Prediction models show that the network dynamics

[15] The code for Erdös–Renyi–Graph node and link simulations using the *igraph* package in R is: for (x in 1:1000){x=x+1; network<-erdos.renyi.game(x,log(1+x/1000),directed=TRUE); print(reciprocity(network));}

[16] See Simonian (2020) for discussion of the importance of time-series cross-validation for financial model building, as well as techniques for bringing time-series cross-validation in line with economic intuition.

EXHIBIT 10

Correlation Network and Hierarchical Clustering for the Funds with Full Track Record: January 2006–December 2016



Note: Mantegna (1999) converted the pairwise correlation coefficient into a topological metric: $d_{i,j} = \sqrt{2(1-\rho)}$.

are largely driven by the momentum, market, value, low beta, and size factors. Incorporating unsupervised learning algorithms into factor models offers investment professionals the ability to enhance their risk analysis and avoid funds, clusters, database categories, and single funds with high centrality scores. Based on cluster analysis and factor exposure, investors can derive asset allocations that are inversely related to importance scores, predict links between funds, and apply predictive regressions to hedge funds at the node level. The approach presented in this article can help investors overcome the confusion associated with database classifications, which differ among vendors. In terms of empirical conclusions, the analysis presented shows that hedge fund relationships change during different parts of the business cycle: Connectedness increases in bad times as liquidity dries up, liquidation probability increases, and lower expected returns become common. To sum up, the methodology presented in this article can enhance investors' arsenal for hedge fund analysis by allowing them to avoid the ambiguity associated with database classifications, improve upon simple connectivity metrics such as correlation, and track the dynamics of hedge fund connectivity as it evolves during the business cycle.

Appendix A

RISK FACTOR STATISTICS

Exhibit A1
Statistics for the Factor Returns: January 2006–December 2016

	BAB	CMA	EM.US	HML	ME($t-1$)	Mkt.RF	MOM	RMW	SMB	YCSpread
Mean	0.0044	0.0016	–0.0029	0.0007	0.0039	0.0050	0.0075	0.0029	0.0002	0.0035
Median	0.0063	–0.0011	–0.0011	0.0006	0.0053	0.0073	0.0092	0.0023	–0.001	0.0025
Maximum	0.1617	0.0572	0.1054	0.047	0.1175	0.1142	0.1295	0.0333	0.0395	1.5604
Minimum	–0.1962	–0.0404	–0.1359	–0.0461	–0.2041	–0.1952	–0.0946	–0.028	–0.0352	–0.8207
Std. Dev.	0.0489	0.0141	0.0433	0.017	0.0478	0.0468	0.0388	0.0111	0.0147	0.2612
Skew	–0.2744	0.5953	–0.0962	0.2353	–0.8297	–0.7728	0.2924	0.0367	–0.1295	1.3898
Kurtosis	5.0363	5.942	3.2482	3.2947	5.3801	5.1487	3.3292	2.9523	2.7317	12.6755

Source: AQR Capital Management, LLC (https://www.aqr.com/Insights/Datasets) and Kenneth R. French's homepage (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html).

Appendix B

FAMILYWISE ERROR RATE

We disclose the familywise error rate, which is the probability of coming to at least one false conclusion in a model and is designed to capture connectedness with p-values < 0.05, which for our model is 36.98% (Fabozzi and López de Prado 2018). Put differently, the probability of a Type I error (false positive) is just over 30.17%. We apply the new p-values (known as Bonferroni correction) by dividing the original acceptance threshold (p-value < 0.05) by the number of tests (new p-value = 0.71% and corresponding t-statistic of ± 2.733) to find the statistical significance of each test. Thus, we controlled for a possible multiple comparisons problem. We select the best model (as suggested by López de Prado 2019) by computing hierarchical clustering and centrality scores.

ACKNOWLEDGMENT

We acknowledge helpful comments from Stephen J. Brown on earlier drafts of this article.

REFERENCES

Adrian, T. 2007. "Measuring Risk in the Hedge Fund Sector." *Current Issues in Economics and Finance* 13 (3): 1–7.

Agarwal, V., and N. Naik. 2000. "Multi-Period Performance Persistence Analysis of Hedge Funds." *Journal of Financial and Quantitative Analysis* 35 (3): 327–342.

Baitinger, E., and T. Maier. 2019. "The (Mis)Behavior of Hedge Fund Strategies: A Network-Based Analysis." *The Journal of Alternative Investments* 22 (1): 57–74.

Baitinger, E., and J. Papenbrock. 2017. "Interconnectedness Risk and Active Portfolio Management." *Journal of Investment Strategies* 6 (2): 63–90.

Bech, M. L., and E. Atalay. 2011. "The Topology of the Federal Funds Market." *Physica A: Statistical Mechanics and its Applications* 389: 5223–5246.

Billio, M., M. Getmansky, A. W. Lo, and L. Pelizzon. 2012. "Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors." *Journal of Financial Economics* 104 (3): 535–559.

Blitz, D. 2018. "Are Hedge Funds on the Other Side of Low-Volatility Trade?" *The Journal of Alternative Investments* 21 (1): 17–26.

Bollen, N. P. B., and R. E. Whaley. 2009. "Hedge Fund Risk Dynamics: Implications for Performance Appraisal." *The Journal of Finance* 64 (2): 985–1035.

Bonacich, P., and P. Lloyd. 2001. "Eigenvector-Like Measures of Centrality for Asymmetric Relations." *Social Networks* 23: 191–201.

Borgatti, S. P. 2005. "Centrality and Network Flow." *Social Networks* 27: 55–71.

Brown, S. J., W. Goetzmann, and J. Park. 2001. "Careers and Survival: Competition and Risk in the Hedge Fund and CTA Industry." *The Journal of Finance* 56 (5): 1869–1886.

Bussiere, M., M. Hoerova, and B. Klaus. 2014. "Communality in Hedge Funds Returns: Driving Factors and Implications." *ECB Working Paper Series* 1658 (3).

Capocci, D., and G. Hübner. 2004. "Analysis of Hedge Fund Performance." *Journal of Empirical Finance* 11 (1): 55–89.

Carhart, M. M. 1997. "On Persistence in Mutual Fund Performance." *The Journal of Finance* 52 (1): 57–82.

Das, S. R. 2016. "Matrix Metrics: Network-Based Systemic Risk Scoring." *The Journal of Alternative Investments* 18 (4): 33–51.

Das, S. R., S. Kim, and D. N. Ostrov. 2019. "Dynamic Systemic Risk: Networks in Data Science." *The Journal of Financial Data Science* 1 (1): 141–158.

Dor, A. B., L. Dynkin, and A. Gould. 2006. "Style Analysis and Classification of Hedge Funds." *The Journal of Alternative Investments* 9 (2): 10–29.

Fama, E. F., and K. R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance* 47 (2): 427–465.

——. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

——. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116: 1–22.

Fama, E. F., and J. D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81 (3): 607–636.

Fabozzi, F. J., and M. López de Prado. 2018. "Being Honest in Backtest Reporting: A Template for Disclosing Multiple Tests." *The Journal of Portfolio Management* 45 (1): 141–147.

Feng, G., S. Giglio, and D. Xiu. 2019. "Taming the Factor Zoo: A Test of New Factors." Fama–Miller working paper, Chicago Booth research article no. 17-04.

Frazzini, A., and L. H. Pedersen. 2014. "Betting against Beta." *Journal of Financial Economics* 111 (1): 1–25.

Fung, W., and D. Hsieh. 1997. "Empirical Characteristics of Dynamic Trading Strategies: The Case of Hedge Funds." *The Review of Financial Studies* 10 (2): 275–302.

——. 2002. "Asset-Based Style Factors for Hedge Funds." *Financial Analysts Journal* 58 (5): 16–27.

Getmansky, M., P. A. Lee, and A. W. Lo. 2015. "Hedge Funds: A Dynamic Industry in Transition." *Annual Review of Financial Economics* 7: 483–577.

Ibbotson, R. G., P. Chen, and K. X. Zhu. 2011. "The ABCs of Hedge Funds: Alphas, Betas, and the Costs." *Financial Analysts Journal* 67 (1): 15–25.

James, G., D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning.* New York: Springer, 2013.

Kazemi, H., Black, K. H., and R. D. Chambers. 2016. *Alternative Investments: CAIA Level II*, 3rd ed. Hoboken: Wiley.

Khandani, A. E., and A. Lo. 2011. "What Happened to the Quants in August 2007? Evidence from Factors and Transactions Data." *Journal of Financial Markets* 14 (1): 1–46.

Kinlaw, W., M. Kritzman, and D. Turkington. 2012. "Toward Determining Systemic Importance." *The Journal of Portfolio Management* 38 (4): 100–111.

Konstantinov, G., A. Chorus, and J. Rebmann. 2020. "A Network and Machine Learning Approach to Factor, Asset, and Blended Allocation." *The Journal of Portfolio Management* 46 (6): 54–71.

Konstantinov, G., and J. Rebmann. 2019. "Different in Nature, Common in Style: View Commonality of Single Hedge Funds and Funds of Hedge Funds." *The Journal of Alternative Investments* (forthcoming).

Konstantinov, G., and M. Rusev. 2020. "The Bond-Equity Fund Relation Using the Fama–French Factors: A Practical Network Approach." *The Journal of Financial Data Science* 2 (1): 22–44.

Kundro, C., and S. Feffer. 2004. "Valuation Issues and Operational Risk in Hedge Funds." *Journal of Financial Transformation* 10: 41–47.

Landes, T., and O. Loistl. 1992. "Complexity Models in Financial Markets." *Applied Stochastic Models and Data Analysis* 8: 209–228.

Liang, B. 2000. "Hedge Funds: The Living and the Dead." *Journal of Financial and Quantitative Analysis* 35 (3): 309–326.

——. 2001. "Hedge Fund Performance: 1990–1999." *Financial Analysts Journal* 57 (1): 11–18.

Loistl, O., and G. S. Konstantinov. 2020. "Interactions and Interconnectedness Shape Financial Market Research." *The Journal of Financial Data Science* 2 (2): 51–63.

Loistl, O., and T. Landes. *The Dynamic Pricing of Financial Assets*. Hamburg: McGraw-Hill. 1989.

López de Prado, M. 2016. "Mathematics and Economics: A Reality Check." *The Journal of Portfolio Management* 43 (1): 5–8.

——. 2019. "A Data Science Solution to the Multiple-Testing Crisis in Financial Research." *The Journal of Financial Data Science* 1 (1): 99–110.

Malkiel, B. G., and A. Saha. 2005. "Hedge Funds: Risk and Return." *Financial Analysis Journal* 61 (6): 80–88.

Mantegna, R. N. 1999. "Hierarchical Structure in Financial Markets." *The European Physical Journal B: Condensed Matter and Complex Systems* 11 (1): 60–70.

Moskowitz, T. J., Y. H. Ooi, and L. H. Pedersen. 2012. "Time Series Momentum." *Journal of Financial Economics* 104 (2): 228–250.

Newman, M. E. J. *Networks: An Introduction*. Oxford: Oxford University Press. 2010.

Pojarliev, M., and R. M. Levich. 2011. "Detecting Crowded Trades." *Financial Analysts Journal* 67 (1): 26–39.

Provost, F., and T. Fawcett. 2013. "Data Science for Business." O'Reilly.

Puliga, M., G. Caldarelli, and S. Battiston. 2014. "Credit Default Swaps Networks and Systemic Risk." *Scientific Reports* 4 (6822).

Raffinot, T. 2018. "Hierarchical Clustering-Based Asset Allocation." *The Journal of Portfolio Management* 44 (2): 89–99.

Sandhu, R. S., T. T. Georgiou, and A. R. Tannenbaum. 2016. "Ricci Curvature: An Economic Indicator for Market Fragility and Systemic Risk." *Science Advances* 2 (5): e1501495.

Schneeweis, T., and R. Spurgin. 1998. "Multifactor Analysis of Hedge Funds, Managed Futures, and Mutual Fund Returns and Risk Characteristics." *The Journal of Alternative Investments* 1 (2): 1–24.

Simonian, J. 2020. "Modular Machine Learning for Model Validation: An Application to the Fundamental Law of Active Management." *The Journal of Financial Data Science* 2 (2): 41–50.

Simonian, J., and C. Wu. 2019a. "Factors in Time: Fine-Tuning Hedge Fund Replication." *The Journal of Portfolio Management* 45 (3): 159–164.

——. 2019b "Minsky vs. Machine: New Foundations for Quant-Macro Investing." *The Journal of Financial Data Science* 1 (2): 94–110.

Simonian, J., C. Wu, D. Itano, and V. Narayanam. 2019. "A Machine Learning Approach to Risk Factors: A Case Study Using the Fama–French–Carhart Model." *The Journal of Financial Data Science* 1 (1): 32–44.

Yang, Z., R. Algesheimer, and C. J. Tessone. 2016. "A Comparative Analysis of Community Detection." *Scientific Reports* 6 (3075).

Disclaimer
This article is for informational purposes only and should not be construed as legal, tax, investment, financial, or other advice. Nothing contained in the article constitutes a solicitation, recommendation, endorsement, or offer by the authors to buy or sell any securities or other financial instruments. The views and opinions expressed in this article are those of the authores only.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Welcome to the Dark Side

Hedge Fund Attrition and Survivorship Bias Over the Period 1994–2001
Gaurav S. Amin and Harry M. Kat
The Journal of Alternative Investments
https://jai.pm-research.com/content/6/1/57

ABSTRACT: *Hedge funds exhibit a high rate of attrition that has increased substantially over time. Lack of size, lack of performance, and an increasingly aggressive attitude of old and new fund managers*

alike appear to be the main factors behind this phenomenon. Although attrition is high, survivorship bias in hedge fund data is modest, which reflects the relatively small difference in performance between surviving and defunct funds. Concentrating on survivors only will overestimate the average hedge fund return by around 2% per annum. For small, young, and leveraged funds, however, the bias can be as high as 4%–5%. The authors also find significant survivorship bias in estimates of the standard deviation, skewness, and kurtosis of individual hedge fund returns. When not corrected for, this will lead investors to seriously overestimate the benefits of hedge funds. Fund-of-funds attrition is much lower. Combined with the small difference in performance between surviving and defunct funds of funds, this yields relatively low survivorship bias estimates for funds of funds.

Do Hedge Funds Hedge?

Clifford S Asness, Robert J Krail, and John M Liew

The Journal of Portfolio Management

https://jpm.pm-research.com/content/28/1/6

ABSTRACT: *Many hedge funds claim to provide significant diversification for traditional portfolios, besides attractive returns. The authors provide empirical evidence regarding the return and diversification benefits of hedge fund investing using the CSFB/Tremont hedge fund indexes over 1994–2000. Like many others, they find that simple regressions of monthly hedge fund excess returns on monthly S&P 500 excess returns seem to support the claims about the benefits of hedge funds. The regressions show only modest market exposure and positive added value. This type of analysis can produce misleading results, however. Many hedge funds hold, to various degrees and combinations, illiquid exchange-traded securities or difficult-to-price over-the-counter securities. For the purposes of monthly reporting, hedge funds often price these securities using either the last available traded prices or estimates of current market prices. These practices can lead to reported monthly hedge fund returns that are not perfectly synchronous with monthly S&P 500 returns. Non-synchronous return data can lead to understated estimates of actual market exposure. When the authors apply standard techniques that account for this problem, they find that hedge funds in the aggregate have significantly more market exposure than simple estimates indicate. Furthermore, after accounting for this increased market exposure, they find that taken as a whole the broad universe of hedge funds does not add value over this period.*

The (Mis)Behavior of Hedge Fund Strategies: *A Network-Based Analysis*

Eduard Baitinger and Thomas Maier

The Journal of Alternative Investments

https://jai.pm-research.com/content/22/1/57

ABSTRACT: *The authors discuss a network-based methodology that models hedge fund strategies across the superordinate–subordinate dimension to gain new insights into their interrelation. This methodology uncovers considerable misbehavior of various hedge fund strategies from the network perspective. Simply speaking, a misbehaving hedge fund strategy has undesired network proximity (similarity) with strategies from other classifications and/or undesired network-based risk properties. The authors conduct extensive static and dynamic (bootstrapping) analyses demonstrating misbehaviors for the full-sample data set. In addition, they demonstrate that numerous network-based behavioral properties of hedge fund strategies can explain future hedge fund returns. This aspect is of significant relevance, as it shows that network-based information has the potential to act as a value-adding warning indicator for funds of hedge funds. Summing up, the authors think that this article provides novel and valuable tools for hedge fund investors, managers, and analysts.*

Derivation of a Dynamic Market Risk Signal Using Kernel PCA and Machine Learning

ALIREZA YAZDANI

ALIREZA YAZDANI is the founder and chief data scientist at Calcolo Analytics LLC in Newton, MA.

KEY FINDINGS

- This article develops a risk signal based on the method of kernel PCA.
- Kernel PCA conveys desirable information on market dynamics.
- Currency machine learning strategies benefit from this kernel PCA feature.

ABSTRACT: *Kernel principal component analysis (PCA) is an extension of the conventional PCA method that employs a kernel transformation whereby hidden patterns in possibly multidimensional data may be detected and extracted more explicitly. In this article, the author applies the method of kernel PCA to a currency prediction case study and derives an aggregate market signal. It is observed that this signal has desirable information-compression properties and may be used as a predictive risk indicator in the return prediction models. Used alongside common drivers of exchange rates, a kernel PCA signal enhances in-sample and out-of-sample risk-adjusted performance across a range of machine learning strategies. In particular, the author observes that a kernel PCA signal remains robust and predictive during volatile market conditions. The kernel PCA signal may be used as a machine learning feature to inform and support data-driven risk management strategies.*

TOPICS: ***Big data/machine learning, derivatives****

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

In finance, data distributions usually change over time owing to idiosyncratic and exogenous factors that are sometimes hard to foresee and quantify. A consequence of this is data with a low signal-to-noise ratio and the strong presence of short-lived and nonlinear patterns, posing serious challenges to many conventional modeling techniques. Machine learning methods, on the other hand, depend less (if at all) on assumptions related to data distributions, which makes them very able to "represent the essential features of the data" (Simonian, López de Prado, and Fabozzi 2018). When rigorously trained and validated, the performance of machine learning models may exceed crude simple methods. This success can be attributed to such methods capturing and benefiting from the information associated with short- and long-term dynamical trends and nonlinear dependencies in data (Li, Turkington, and Yazdani 2020). These properties make machine learning particularly attractive from a finance application standpoint.

One useful application of machine learning is concerned with developing

indexlike signals. Generally speaking, an index is a broad indicator of the movements and performance of financial markets. It represents common or systematic patterns in specific segments of the market and may be used as a performance benchmark for investment portfolios and passive investing strategies or to reflect risk premiums associated with market constituents. Notable examples of indexes are the S&P 500 Composite Stock Price Index and Dow Jones Industrial Average in the equities market and the US Dollar Strength Index (USDX) (ICE 2015, 2018) in the currency market.

Indexes typically use simple weighting methods as the basis of calculation, which makes them easy to both calculate and follow and investable. For example, the USDX uses a bilateral trade-weighted geometric average of six major currencies: euro (EUR), Japanese yen (JPY), British pound (GBP), Canadian dollar (CAD), Swedish krona (SEK), and Swiss franc (CHF). The USDX tracks the strength of the US dollar such that it increases when the US dollar gains strength (value) compared to other currencies. However, the USDX is by design restricted to only a few currencies with fixed weighting factors and has not been adjusted since the introduction of the euro. As such, it may not adequately reflect the impacts of short-lived and regional events across the broader currency markets such as the G10 or emerging markets (ICE 2018).

Subsequently, using data-driven methods to develop indexlike dynamic market signals may improve overall information-compression properties and help better design and develop risk and return strategies for currency portfolios. Such signals are analogous to dynamic indexes (Lo 2016) because they are transparent (reproducible) and systematic (data driven), although they may not always be directly investable given the underlying construction mechanism used. In particular, because of this data-driven construct, few discretionary investment skills are needed to apply them on an ongoing basis. As such, sophisticated investors may largely benefit from using them in practice as part of their investment risk management enhancement strategy.

In what follows, we construct a dynamic currency market risk signal using a method known as kernel principal component analysis (PCA). This method offers a nonlinear extension of PCA by employing a nonlinear kernel transformation. We first apply the kernel PCA method to the same set of currencies used in the USDX index and compare the outcomes. We then extend this notion to the set of G10 currencies and carry out experiments to evaluate its properties related to compression and reflection of the currency market movements, risk, and volatility estimations. In theory, kernel PCA is closely related to another machine learning method known as *spectral clustering* (Ng, Jordan, and Weiss 2002; Bengio, Vincent, and Paiement 2003). Spectral clustering has been demonstrated recently as a powerful data-driven basis for regime detection in macro investing (Simonian and Wu 2019). We are similarly motivated to explore kernel PCA in investment risk management as a useful extension of the PCA methodology that is widely used in quantitative finance applications.

Through examples, we demonstrate that the kernel PCA method applied on the spot returns covariance matrix of G10 currencies can be used alongside other known drivers of the foreign exchange market to predict returns and risk. In particular, we observe that such a kernel PCA feature can boost the performance of currency return prediction (machine learning) models above and beyond simpler strategies and other aggregated market risk factors. We attribute these observations to useful market information conveyance properties of kernel PCA and its ability to track short- and mid-term market moves; it is particularly useful as a predictor of extremes when used in a machine learning framework.

KERNEL PCA METHODOLOGY

PCA is a statistical technique that converts a matrix of covariates (i.e., possibly correlated observations) into a matrix of uncorrelated variables called *principal components* with the property that the first few principal components reflect the directions along which original data vary the most. Selecting the first (or first few) components is therefore an effective way to reduce the dimensionality of data while preserving the most useful information. Conventional principal components can be obtained from the eigenvectors of the covariance matrix of covariates. Kernel PCA (Schölkopf, Smola, and Müller 1998), viewed as a nonlinear expansion of PCA, expands the feature space by first using a nonlinear kernel transformation and then applying PCA in this transformed space. The destination feature space of this transformation may have higher dimensions than

EXHIBIT 1

Summary Statistics of the First Principal Components of Different Kernels

Method	Kernel	Kernel Parameter(s)	Explained Variance	Median	Mean	Kurtosis	Skewness	5th Percentile	95th Percentile
PCA	NA	NA	65.8%	0.02	0.00	–0.44	–0.47	–0.75	0.51
KPCA_G	Gaussian	σ = 0.001	72.7%	0.05	0.01	–0.49	–0.52	–1.99	1.36
KPCA_L	Laplacian	σ = 0.001	64.9%	0.06	0.01	–0.89	–0.43	–1.85	1.32
KPCA_P1	Polynomial	d = 1, a = 1, b = 0	65.8%	0.05	0.01	–0.49	–0.52	–1.99	1.36
KPCA_P3	Polynomial	d = 3, a = 1, b = 0	67.0%	0.27	0.00	–0.41	–0.73	–2.04	1.26
KPCA_P5	Polynomial	d = 5, a = 1, b = 0	70.4%	0.40	0.00	–0.38	–0.80	–2.05	1.19

original input space, thus allowing better separation of data using a hyperplane.

The underlying mechanism of kernel PCA is simple yet powerful. Given a set of (zero-centered) data $x_k \in R^n$ and $k = 1, 2, \ldots, m$, the PCA method finds eigenvectors ν by solving the equation $\lambda\nu = C\nu$, where $C = \frac{1}{m}\Sigma_{j=1}^{m} x_j x_j^T$ is the diagonalized covariance matrix of observations. Now, the idea is to map (linearly inseparable) data into a higher-dimensional space where different points can be separated more easily (i.e., linearly). For such typically nonlinear transformation ϕ: $R^n \rightarrow R^d$ we can write $\hat{C} = \frac{1}{C}\Sigma_{j=1}^{n}\phi(x_j)\phi(x_j)^T$ and compute the eigenvectors of the centered kernel matrix $K_{ij} = \langle \phi(x_i) | \phi(x_j) \rangle = \phi(x_i)\phi(x_j)^T$. The conventional PCA method may be obtained as a special case of kernel PCA by choosing the linear identity kernel function $\phi(x) = x$.

Different kernels may be used in the process to detect nonlinearities, although the choice of one such kernel may not be known in advance. Examples of commonly used kernels include the Gaussian kernel

$$k(x, y) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right),$$

Laplace kernel $k(x, y) = \exp(-\sigma||x - y||)$, and degree-$d$ polynomial kernel $k(x, y) = (a{*}x^T y + b)^d$. Different kernels reshape input data in a destination space differently and result in different decision boundaries. In general, a Gaussian may serve well many applications in which little prior knowledge about data is available, although one such choice of kernel and its parameter(s) may later be adjusted as the nature of the data and problem demand. For theoretical details and some applications of kernel PCA, see Hastie et al. (2008), Bishop (2006), and Karatzoglou et al. (2004).

KERNEL PCA AS A DYNAMIC MARKET INDEX

We now compare some properties of kernel PCA index against those of the USDX index. To construct the USDX index, we use a weighted average of six major currency spot rates sampled at the end of each month from March 1998 to September 2019. The original set of currencies included in USDX is (ICE 2015) EUR, JPY, GBP, CAD, SEK, and CHF. The USDX weighted average formula is given by

$$\begin{aligned} USDX_t = 50.1434 * (&EURUSD_t^{-0.576} * USDJPY_t^{0.136} \\ &* GBPUSD_t^{-0.119} * USDCAD_t^{0.091} \\ &* USDSEK_t^{0.042} * USDCHF_t^{0.036}) \end{aligned}$$

where the subscript t indicates the measurement frequency. It is important to note that weight factors for leading multipliers and exponents of USDX are currently fixed (since the introduction of EUR) with no ongoing change. Next, we use different kernels to construct a few kernel PCA signals that are applied to the same basket of currencies used in USDX.

Exhibit 1 reports some descriptive statistics and the variance explained by the first principal component from each method. The kernel PCA using Gaussian kernel seems to be capable of capturing the largest variance (circa 72.7%) from the set of six currencies. This is followed by a degree-5 polynomial kernel, which seems excessive and dubious owing to the potential risk of in-sample overfitting (we will not pursue it further).

Exhibit 2 shows the reconstructed time series of USDX, compared with values from the first principal component of a simple PCA and different kernel PCA methods centered at zero. A visual inspection of results

EXHIBIT 2

Comparison of PCA, Kernel PCA, and USDX Index

suggests that different kernel PCAs, and to some extent USDX, convey similar yet distinct information historically, whereas some divergence can be observed in recent years (e.g., after 2015). Conventional PCA, on the other hand, yields a much smoother pattern, closely hovering around the mean and exhibiting fewer extremes and peaks.

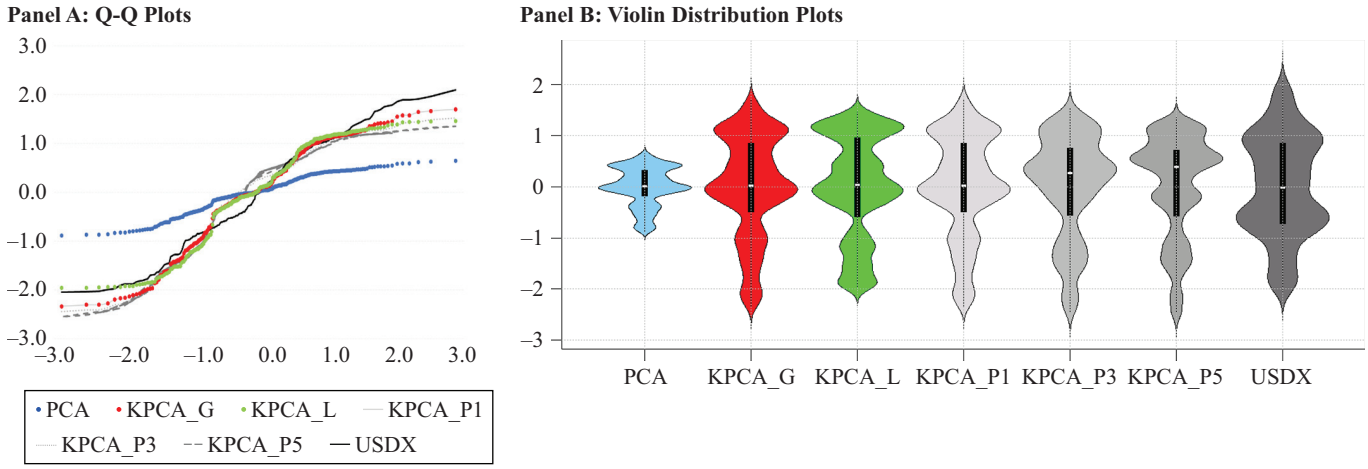
Exhibit 3 depicts the Q-Q plot (Panel A) and violin distribution plots (Panel B) of the first principal component values from each kernel method and compares them against the USDX. The Gaussian kernel with a scale parameter of $\sigma = 0.001$ captures most variations from the six currencies, followed by a degree-5 polynomial kernel. For the studied data, different kernel PCA methods more greatly resemble each other in terms of the measures of variation and distributional properties, and somewhat resemble a more symmetrically distributed USDX, than a simple PCA.

It is important to note that these kernel methods and their parameters are chosen somewhat arbitrarily, and we may evaluate their performance as in sample only. A natural question to ask is how kernel PCA methodology can be used to reflect and anticipate future market moves beyond explaining in hindsight what happened in the past. Naturally, this is a question best answered via empirical analysis, which in turn depends on study objectives partly defined by data specifications, response sensitivity, and required predictive accuracy in practice. In what follows, we demonstrate how the kernel PCA method may be used in practice as a factor (or *feature*, in machine learning jargon) to inform currency returns prediction in an investment portfolio setting.

KERNEL PCA AS A MACHINE LEARNING FEATURE

We would like to study whether and how a kernel PCA feature contributes to machine learning models' ability to learn and use patterns in data both inside and outside of a training sample. Before we proceed, some modification is required in the process of constructing a kernel PCA feature. First, we expand our universe

EXHIBIT 3
Comparison of Kernel PCA, PCA, and USDX Q-Q Plots



of currencies to the larger and more liquid basket of G10 (developed market) currencies to broaden the scope of our analysis by incorporating a universe of possible arbitrage opportunities larger than just six currencies included in the USDX. The G10 currencies include the Australian dollar, CAD, CHF, EUR, GBP, JPY, Norwegian krone (NOK), New Zealand dollar, SEK, and USD.

Second, unlike the previous section in which spot rates were used for exploratory and comparison purposes only, we apply kernel PCA to pairwise G10 currency returns instead. Following Li, Turkington, and Yazdani (2020), we form a dataset of monthly returns for each of the exchange rate pairs from the G10 currencies, resulting in 45 non-overlapping pairs. Using a returns covariance matrix to build kernel PCA helps tease out relatedness and volatilities in currency returns and complies with an important consideration in predictive modeling: avoiding features that are dependent on nonstationary data that may become unstable over time. In addition, we apply one order lag to the first principal component values obtained from applying kernel PCA to reduce the risk of data leakage and look-ahead bias.

For the portfolio construction phase, we identify each month 27 out of 45 currency pairs with the largest prediction magnitude and assign long or short positions, depending on each prediction's directional sign. The selection of 27 out of the 45 pairs helps diversify the portfolio by avoiding significant domination by one single currency, similar to the conventional top-three/bottom-three long–short portfolio construction. Similarly, we structure the prediction problem as a panel regression in which we use point-in-time information available about given currency pairs to predict on-month forward returns, as investable exposures in the currency market. Factors used in this prediction consist of a parsimonious set of known drivers of currency returns, including

- *carry*—the short-term interest rate differential between countries (Burnside, Eichenbaum, and Rebelo 2011)
- *trend*—the trailing one-year spot return for a currency pair adjusted for one-month reversals (Burnside, Eichenbaum, and Rebelo 2011)
- *equity differential*—the trailing one-year equity return differential between countries (Turkington and Yazdani 2020)
- *valuation*—the trailing five-year spot return adjusted for trailing five-year inflation differential, to indicate fair value for currency pairs (Turkington and Yazdani 2020)
- the kernel PCA feature described previously, derived using a Gaussian kernel method

In addition to employing these factors as input features in regression models, we use them individually to construct portfolios that represent single-factor investment strategies. These strategies do not use any predictive model but rather are based on the long or

EXHIBIT 4

Strategy Performance Excluding Kernel PCA (training sample)

	Carry	Trend	Valuation	Equity Differential	OLS	RF	GBM	ANN
Return	3.58%	0.73%	2.32%	2.49%	3.61%	3.91%	4.16%	3.79%
Risk	6.52%	6.88%	6.19%	4.98%	5.30%	6.21%	4.65%	5.74%
Ratio	0.55	0.11	0.37	0.50	0.68	0.63	0.90	0.66
Skewness	–0.71	–0.04	0.98	–0.33	–0.60	0.51	0.04	–1.18
Kurtosis	1.35	3.50	4.62	0.97	2.12	3.02	3.06	3.59
Hit Rate	63%	53%	52%	58%	61%	57%	60%	69%

short investment signal constructed from that factor. Regression models we use to predict currency returns include ordinary least squares (OLS), random forest (RF), gradient boosting machine (GBM), and a shallow regularized artificial neural network (ANN). The choice of these models is based on the premise of diversity in model characteristics (e.g., nonlinear, tree based) and learning capabilities. For more details on these machine learning models, see Hastie, Tibshirani, and Friedman (2008); Kuhn and Johnson (2013); and Goodfellow, Bengio, and Courville (2016). Finally, for the first pass of model training and evaluation, we split data into a training sample spanning January 1990 to December 2013 and a test sample spanning January 2014 to March 2019, held out of model training and used only for final performance evaluation.

TRAINING SAMPLE PERFORMANCE

We would like to know whether, to what extent, and how a kernel PCA feature adds to the predictive power of different machine learning strategies. In particular, we are interested in knowing if the risk-adjusted performance of these models exceeds that of the single-factor currency strategies and a benchmark linear (OLS) regression model. To this end, we first train machine learning models without a kernel PCA feature, using only carry, trend, valuation, and equity differential factors, and evaluate and compare performance. We then add a kernel PCA feature, retrain, and report performance. As a part of machine learning model training, we use a 10-fold time-block cross validation (to preserve time ordering and serial dependencies) and tune hyperparameters to minimize root-mean-square error.

Exhibit 4 shows the training sample (January 1990–December 2013) performance of different strategies, including portfolios formed on only single factors, as well as models that use the four predictors of carry, trend, valuation, and equity differential factors but exclude the kernel PCA feature. The performance metrics reported here include the first few moments of the strategy returns, including annualized mean (return) and variance (risk), return skewness and kurtosis, the (information) ratio of return to risk, and the percentage of months with a positive return (hit rate).

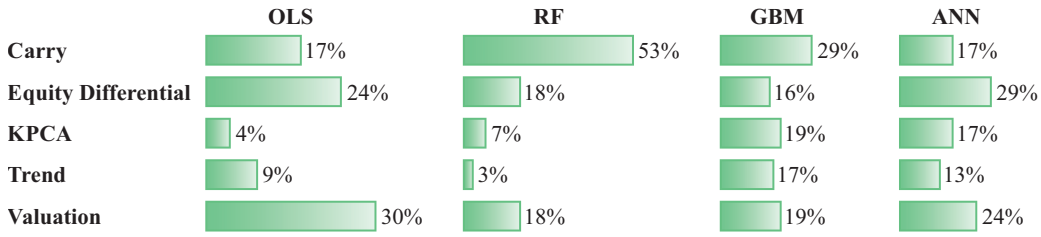
Multifactor models, particularly GBM, outperform traditional single-factor strategies in terms of annualized risk-adjusted return (information ratio). The GBM model has the best training sample performance with the highest annualized return and lowest risk levels. The RF model generates the second highest annualized return but also a relatively high risk level. The ANN has the highest percentage of months with a positive return, and the OLS model performs decently in comparison. Notably, the carry strategy generates the best results among the single factors, including the highest ratio and hit rate, although it becomes a highly risky strategy in the same time period. Overall, these results are not surprising, given the capability of the machine learning models in processing multiple features and finding signals that are not accessible to simpler strategies; however, with only four factors to process, data-hungry machine learning models may not achieve their maximum performance.

Exhibit 5 shows training sample performance of these strategies but with a Gaussian kernel PCA feature added to the list of predictors in multifactor models. The single-factor strategy results, independent from kernel PCA and hence identical to Exhibit 4 results, are included here for the sake of completeness. Adding a kernel PCA feature makes a positive contribution to the performance of machine learning strategies by generating higher returns and lower risk (except for OLS). In addition, the distributional properties do not exhibit

EXHIBIT 5
Strategy Performance Including Kernel PCA (training sample)

	Carry	Trend	Valuation	Equity Differential	OLS	RF	GBM	ANN
Return	3.58%	0.73%	2.32%	2.49%	3.68%	4.72%	4.31%	4.59%
Risk	6.52%	6.88%	6.19%	4.98%	5.33%	6.16%	4.59%	4.82%
Ratio	0.55	0.11	0.37	0.50	0.69	0.77	0.94	0.95
Skewness	–0.71	–0.04	0.98	–0.33	–0.55	0.53	–0.07	–0.63
Kurtosis	1.35	3.50	4.62	0.97	2.38	3.64	2.54	2.29
Hit Rate	63%	53%	52%	58%	63%	61%	63%	65%

EXHIBIT 6
Relative Contribution by Each Factor per Model



significant drift, and hit rates generally improve with a kernel PCA feature. We will explore the likely causes of this observation in more detail later.

Exhibit 6 shows the percentage of overall contribution by each factor to predictions in the training sample, as a basis of feature importance. Importance is generally calculated based on the reduction in squared error achieved as a result of adding a feature in machine learning models (Kuhn 2008). The contributions to prediction in OLS and RF are dominated by valuation and carry factors, respectively, whereas GBM and ANN models benefit from contributions from a diverse set of factors, kernel PCA included.

A brief feature correlation analysis indicates no major correlation among predictors. Exhibit 7 shows the regression panel's pool-level factor correlations.

At an individual-currency-pair level, correlation between factors varies but typically in the same range as in the pool. In particular, the pair-level correlation between kernel PCA and other factors is as follows: with carry, between –11% and 7% with a mean of –2%; with equity differential, between –12% and 5% with a mean of –2%; with valuation, between –14% and 19% with a mean of 3%; and with trend, between –22% and 26% with a mean of 1%. Finally, a standard test of

EXHIBIT 7
Feature Correlations (training sample)

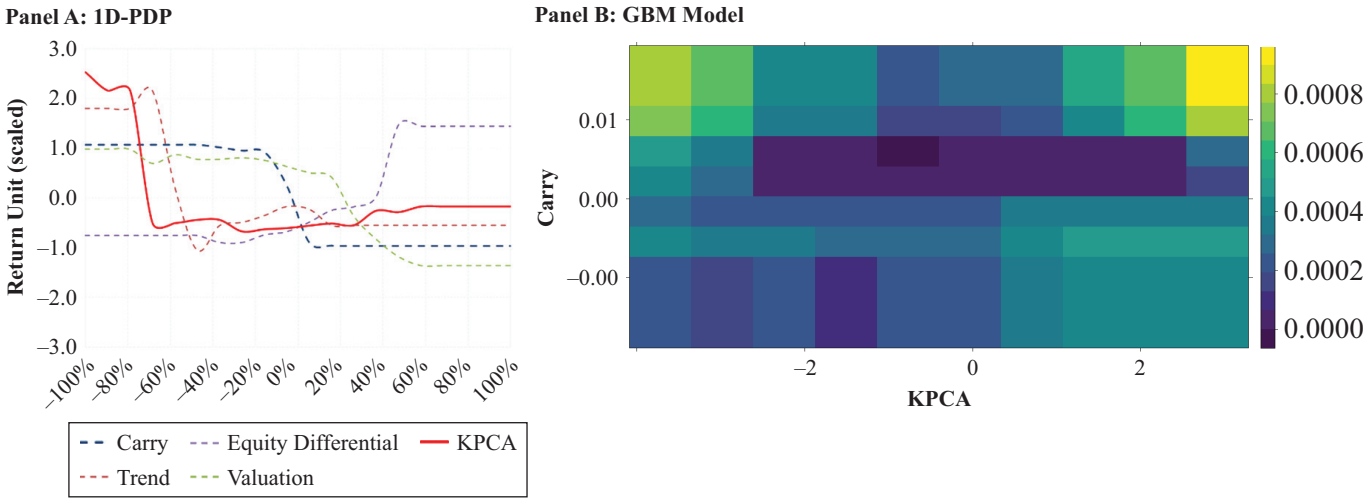
	KPCA	Carry	Equity Differential	Valuation
Carry	–2%			
Equity Differential	–2%	–1%		
Valuation	2%	–21%	–11%	
Trend	1%	–5%	–5%	31%

multicollinearity applied to the OLS model yields a maximum variance inflation factor of around 1.16, very close to a theoretical lower bound of 1 and far from the cutoff value of 5 sometimes used as the threshold to detect large pairwise multicollinearity in regression models (James et al. 2017). Subsequently, we safely assume that the features used convey relevant and independent information for predictions, none is largely explained by other features, and they may be used in conjunction.

To better understand how the previously mentioned machine learning models detect nonlinear effects and variable interactions in data, we employ a method known as *partial dependency* (Friedman 2001) in its one- and two-dimensional forms. A one-dimensional partial

EXHIBIT 8

1D-PDP and Carry-Kernel PCA Interaction in GBM Model



dependency quantifies the marginal impact of one feature on model predictions, and a two-dimensional partial dependency shows the joint impact from two variables. Exhibit 8 depicts how the GBM model uses information from different features in a nonlinear and interactive fashion. As seen in the one-dimensional partial dependency plot (1D-PDP) in Exhibit 8, Panel A, predicted returns respond nonlinearly to different features as enabled by the ability of GBM model to learn nonlinear patterns in data (it may be worth noting that partial dependency always produces a line in the OLS case). For instance, large negative (i.e., bottom quintile) values of kernel PCA features are highly predictive and generate large positive returns, although such returns are largely diminished after kernel PCA values pass a certain threshold. In particular, there is similarity in the directionality and nonlinear patterns of how GBM uses kernel PCA and trend features that may in part be caused by the slightly higher than average pair-level correlations observed earlier. Similarly, GBM processes other factors in a nonlinear way but, perhaps, in a reverse direction (e.g., equity differential). As such, the partial dependency is analogous to a model-dependent nonlinear extension of the values and signs of coefficients in a simple linear regression. Note that here values in both the horizontal and vertical axes are rescaled for comparison and visualization purposes, but they faithfully preserve the directionality and various contribution levels by each factor. In other words, the returns predicted in raw scales by kernel PCA are also generally higher than other factors.

The heatmap in Exhibit 8 is an example of two-dimensional partial dependency, specifically showing interaction between kernel PCA and carry as detected by GBM. Despite overall mixed effects, higher returns are generated (top right corner) when the highly positive carry effect overlaps the highly positive kernel PCA. For a customized approach to unraveling and interpretating nonlinear and interaction effects in machine learning, see Li, Turkington, and Yazdani (2020).

TEST SAMPLE PERFORMANCE

To have a more realistic evaluation of model performance, we make predictions on a holdout test sample and evaluate portfolios. Test sample data span January 2014 to March 2019 and have not been used in any stage of model training. First, we look at the test sample performance of single-factor strategies and models excluding the kernel PCA feature, shown in Exhibit 9. Notably, a single-factor equity differential strategy shows very strong results in recent years; its risk-adjusted performance exceeds that of model. This strong performance may be attributed to investors chasing higher returns from local markets marginally driving currency values, as argued by Turkington and Yazdani (2020). The other single-factor strategies of carry, trend, and valuation seem to have faltered in recent years owing to the reduced opportunity associated with such factors. Another interesting observation is that multifactor models on average generate a higher

EXHIBIT 9
Strategy Performance Excluding Kernel PCA (test sample)

	Carry	Trend	Valuation	Equity Differential	OLS	RF	GBM	ANN
Return	0.56%	–1.72%	–0.19%	3.57%	2.22%	1.40%	2.24%	1.86%
Risk	4.94%	4.40%	4.04%	4.47%	4.28%	2.58%	3.34%	4.17%
Ratio	0.11	–0.39	–0.05	0.80	0.52	0.54	0.67	0.45
Skewness	–0.52	–0.28	0.07	0.62	0.13	2.00	0.55	–0.28
Kurtosis	1.00	–0.21	–0.43	0.29	1.13	9.42	1.96	0.62
Hit Rate	51%	51%	46%	52%	60%	51%	57%	59%

EXHIBIT 10
Strategy Performance Including Kernel PCA (test data)

	Carry	Trend	Valuation	Equity Differential	OLS	RF	GBM	ANN
Return	0.56%	–1.72%	–0.19%	3.57%	2.26%	1.67%	2.88%	2.73%
Risk	4.94%	4.40%	4.04%	4.47%	4.23%	2.80%	3.27%	4.18%
Ratio	0.11	–0.39	–0.05	0.80	0.54	0.60	0.88	0.65
Skewness	–0.52	–0.28	0.07	0.62	0.06	0.71	0.54	0.12
Kurtosis	1.00	–0.21	–0.43	0.29	0.99	1.92	2.18	1.21
Hit Rate	51%	51%	46%	52%	59%	57%	59%	62%

percentage of positive returns, as indicated by hit rates. Similar to the training sample, GBM performs best among multifactor models by generating comparatively higher returns and lower risk. Finally, we observe that the RF model returns have a very high kurtosis, indicating prediction of extreme returns.

Exhibit 10 depicts the test sample performance of the same models when they include the kernel PCA feature and shows the performance of single-factor strategies for the sake of completeness. There is a clear boost in test sample performance as a result of using the kernel PCA feature. In particular, GBM and ANN largely benefit from this additional feature in terms of predicting higher returns, lower risk, and an over 20% increase in the annualized return-to-risk ratio. Exhibit 11 shows a full-sample cumulative return series for all strategies. Pulling things together, multifactor strategies seem to benefit from the kernel PCA feature significantly, with their performance improving both inside and outside the training sample. The likely explanation is that machine learning models achieve this through detecting market movements and better timing market risk, thanks to the information conveyed by the kernel PCA feature. We further explore this hypothesis in the next section.

KERNEL PCA AS A RISK SIGNAL

As mentioned in an earlier section, USDX has a straightforward interpretation as a measure of USD's relative strength. Perhaps kernel PCA may be viewed as an extension of this notion, albeit using the G10 currency basket with a potentially more dynamic (yet sophisticated) construct. It would be more useful if the kernel PCA factor could be used as a market indicator conveying information associated with extreme recent market movements and systemic risk. To establish this, we conduct two additional experiments by building machine learning strategies that use two alternative risk factors; we compare their performance against strategies that are built using the kernel PCA feature. These two alternative factors, described in the following, are the generalized autoregressive conditional heteroskedastic (GARCH) volatility estimate and the G10 Currency Turbulence Index.

The GARCH model is a volatility modeling technique that estimates the variance of a financial returns series using the historical variance of the series and error terms from an autoregressive model of estimating (squared) returns (Bollerslev 1986). GARCH models are widely used to assess conditional risk for assets that

EXHIBIT 11

Cumulative Returns Including Kernel PCA (full sample)

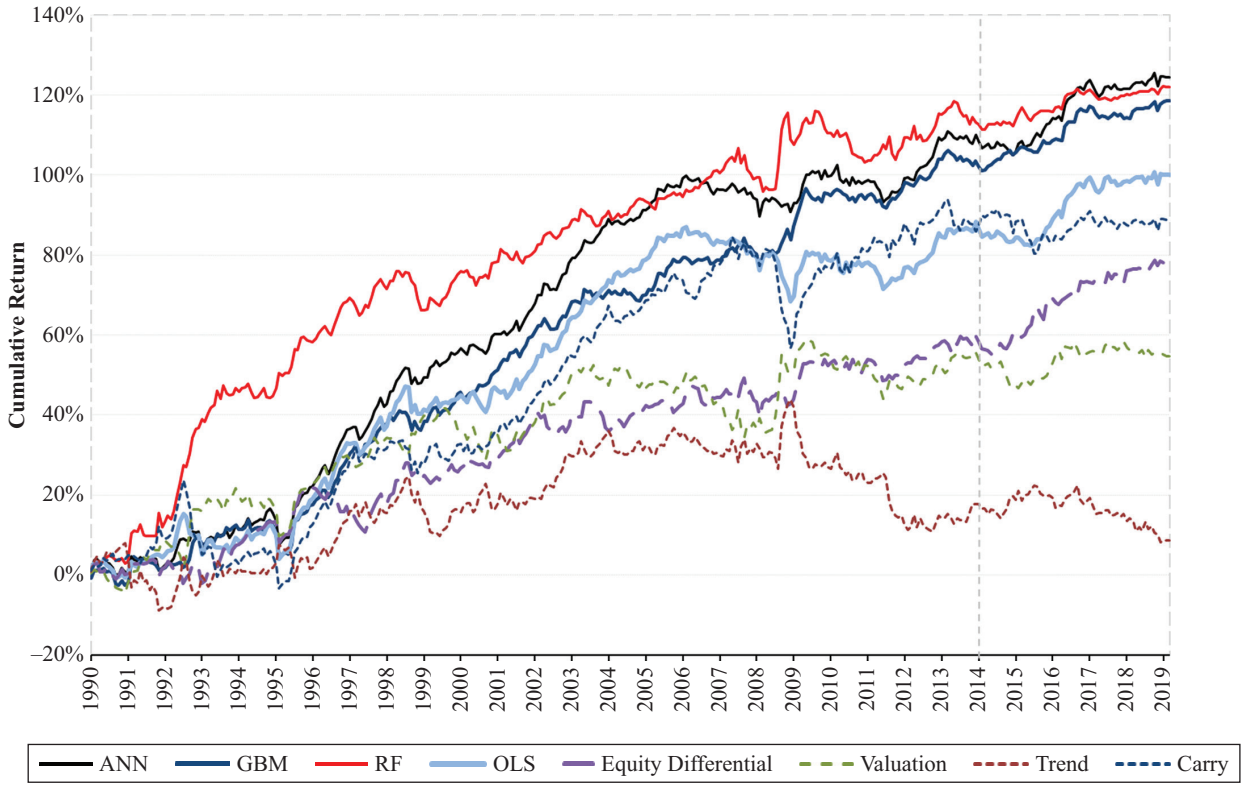


exhibit time-varying and clustered volatility. We use a GARCH(1,1) with normal distribution model with a volatility term order of one and an (error term) autoregressive order of one, which is used most commonly in practice owing to its computational simplicity and adequacy. We estimate a univariate GARCH model for each of 45 pairs and use it as an individual volatility feature in machine learning models. Additionally, we use the G10 Currency Turbulence Index (Kritzman and Li 2010) as an alternative machine learning feature. Turbulence is a multivariate score, based on the Mahalanobis distance, representing an aggregate market risk profile. The originators of the Turbulence Index state that it is a persistent indicator that can be used to inform and anticipate unusual market moves, correlation breakdowns, and reverse factor behavior (e.g., in carry trade).

In our first experiment, similar to earlier sections, we develop strategies using historical data between January 1990 and December 2013 and holding out data between January 2014 and March 2019 as the test sample for the purpose of making predictions on unseen data. Here, we use the kernel PCA feature, GARCH factor, and Turbulence Index alternatively as a *risk signal* to train multifactor models. Exhibit 12 shows the training sample return-to-risk ratio of selected portfolios, including the top two single-factor strategies of carry and equity differential, as well as two models including the standard linear regression OLS model and GBM machine learning model. The results for RF and ANN are similar to but slightly weaker than those for GBM. Generally speaking, in-sample model performance improves as we move from using kernel PCA to GARCH and to the Turbulence Index.

On the other hand, it is crucial to study these alternative strategies on previously unseen test data as a more realistic evaluation of performance. Exhibit 13 shows the performance of the same strategies on test data. The test sample performance shows a different rank ordering of model performance: GBM using kernel PCA marginally outperforms GARCH, which in turn outperforms the Turbulence Index. Overall, the OLS linear model seems to benefit more from the GARCH signal, reflecting the volatility of individual currency pairs. All GBM models

EXHIBIT 12
Strategy Return-to Risk Ratio (training data)

	Carry	Equity Differential	OLS	GBM
KPCA	0.55	0.50	0.69	0.94
GARCH	0.55	0.50	0.77	0.94
TURBULENCE	0.55	0.50	0.92	1.01

EXHIBIT 13
Strategy Return-to-Risk Ratio (test data)

	Carry	Equity Differential	OLS	GBM
KPCA	0.11	0.80	0.54	0.88
GARCH	0.11	0.80	0.65	0.87
TURBULENCE	0.11	0.80	0.36	0.80

outperform single-factor strategies and the OLS model in both training and test data. We attribute these observations to the ability of machine learning methods to identify and benefit from intricate nonlinear dynamics conveyed by the kernel PCA factor.

In our second experiment, we change model training and prediction windows to different time spans. Here, we train models on historical data between January 1990 and December 2007 and hold out data between January 2008 and December 2011 for testing. Similar to our first experiment, we use kernel PCA, GARCH, and the Turbulence Index as alternative model features. We then make predictions out of sample and measure and compare portfolio performances on both samples. With this approach, we basically train models on a (shorter) portion of the historical data used earlier but evaluate their performance using a test sample associated with the global financial crisis, which affected the currency markets starting in late 2007 and continuing for the next few years. This particular selection of test sample may pose a serious challenge to the ability of many models to generalize and carry over their learning from a relatively calm training period to an extremely unusual and volatile out-of-sample period (i.e., 2008–2011). It is worth remembering that, all else being equal, any differences in the test sample predictions of these strategies are primarily driven by the use of alternative features, namely, kernel PCA, GARCH, and the Turbulence Index.

Exhibit 14 shows the training sample performance (annualized return-to-risk ratio) of the selected strategies

EXHIBIT 14
Strategy Return-to Risk Ratio (training data, January 1990–December 2007)

	Carry	Equity Differential	OLS	GBM
KPCA	0.74	0.48	0.80	1.39
GARCH	0.74	0.48	1.02	1.27
TURBULENCE	0.74	0.48	0.98	1.33

EXHIBIT 15
Strategy Return-to Risk Ratio (test data, January 2008–December 2011)

	Carry	Equity Differential	OLS	GBM
KPCA	0.17	0.34	–0.46	0.45
GARCH	0.17	0.34	–0.57	–0.03
TURBULENCE	0.17	0.34	–0.30	0.36

also discussed in the first experiment (i.e., Exhibit 12). The rank ordering of performance remains the same as before: Machine learning models outperform OLS, which in turn outperforms single-factor strategies. The OLS model benefits most from the GARCH volatility signal, whereas the GBM model clearly benefits from the kernel PCA feature. Notably, the GBM model generates a return-to-risk ratio nearly twice as high as that of the carry strategy, which is regarded as a historically profitable strategy in the years leading to the global financial crisis. The picture, however, looks very different for the period of January 2008 through December 2011, associated with the crisis (Exhibit 15). The carry effect drastically disappears, and the equity differential strategy deteriorates. Worst of all, the OLS strategy is crushed in this period, a clear indication of the linear model's salient property of fitting to historical mean, which results in a lack of flexibility to time the market and diversify risk. The GBM model, on the other hand, benefits more from using aggregate market risk signals (primarily kernel PCA but also turbulence) during this volatile period, above and beyond using individual currency pairs' volatility measure of GARCH. This clearly indicates that machine learning strategies benefiting from strong and inherently dynamic market risk signals may outperform simple strategies by identifying temporal nonlinear and interactive patterns in data.

CONCLUSION

We explore how a kernel PCA, a nonlinear extension of the conventional PCA, can be applied to a basket of currencies to develop a data-driven method for a currency market aggregate risk signal. We argue that, despite their apparent sophistication, data-driven methods such as kernel PCA exhibit strong dynamical properties, enabling them to capture nonlinear and interactive relations in the market. We employ machine learning currency prediction strategies that use kernel PCA as an input feature and show that they may outperform a number of traditional single-factor and simpler modeling strategies, both in sample and out of sample. We also show that the risk-adjusted performance of machine learning models using the kernel PCA feature is largely robust under different scenarios. These scenarios include comparison against established risk indexes (including the currency Turbulence Index as an aggregate risk score) and the univariate GARCH volatility estimate for individual currency pairs, as well as exposure to different market regimes such as the global financial crisis through variations in the training and test windows of the machine learning algorithms. We highlight two main reasons for the effectiveness of the described data-driven investment strategies: (1) Compressed dynamic market information conveyed through a kernel PCA signal and (2) the ability of machine learning methods to effectively process information and detect signals that may not be accessible to traditional single-factor strategies or simple linear models. In summary, we argue that carefully calibrated machine learning methods can make important contributions in a variety of settings and scenarios concerning investment portfolios' risk-and-return strategies and overcome historically challenging problems that are hard to cope with using traditional strategies.

ACKNOWLEDGMENT

I thank David Turkington and Joseph Simonian for their feedback.

REFERENCES

Bengio, Y., P. Vincent, and J. Paiement. 2003. "Spectral Clustering and Kernel PCA Are Learning Eigenfunctions." Technical report 2003s-19, CIRANO, https://cirano.qc.ca/files/publications/2003s-19.pdf.

Bishop, C. M. *Pattern Recognition and Machine Learning*. New York: Springer. 2006.

Bollerslev, T. 1986. "Generalized Autoregressive Conditional Heteroscedasticity." *Journal of Econometrics* 31: 307–327.

Burnside, C., M. Eichenbaum, and S. Rebelo. 2011. "Carry Trade and Momentum in Currency Markets." *Annual Review of Financial Economics* 3 (1): 511–535.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5): 1189–1232.

Goodfellow, I., Y. Bengio, and A. Courville. *Deep Learning*. Cambridge, MA: MIT Press. 2016.

Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer. 2008.

ICE. "US Dollar Index Contracts." 2015, https://www.theice.com/publicdocs/futures_us/ICE_Dollar_Index_FAQ.pdf.

——. 2018. "ICE FX Indexes (DXY, ICELX, ICEEX, ICEJX)." https://www.theice.com/publicdocs/data/ICE_FX_Indexes_Methodology.pdf.

James, G., D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*, 8th ed. New York: Springer. 2017.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis. 2004. "Kernlab—An S4 Package for Kernel Methods in R." *Journal of Statistical Software* 11 (9).

Kritzman, M., and Y. Li. 2010. "Skulls, Financial Turbulence, and Risk Management." *Financial Analysts Journal* 66 (5): 30–41.

Kuhn, M. 2008. "Building Predictive Models in R Using the Caret Package." *Journal of Statistical Software* 28 (5).

Kuhn, M., and K. Johnson. *Applied Predictive Modeling*. New York: Springer. 2013.

Li, Y., D. Turkington, and A. Yazdani. 2020. "Beyond the Black Box: An Intuitive Approach to Investment Prediction with Machine Learning." *The Journal of Financial Data Science* 2 (1): 61–75.

Lo, A. 2016. "What Is an Index?" *The Journal of Portfolio Management* 42 (2): 21–36.

Ng, A., M. Jordan, and Y. Weiss. 2002. "On Spectral Clustering: Analysis and an Algorithm." *Advances in Neural Information Processing Systems*.

Schölkopf, B., A. Smola, and K. Müller. 1998. "Nonlinear Component Analysis as a Kernel Eigenvalue Problem." *Neural Computation* 10 (5): 1299–1319.

Simonian, J., M. López de Prado, and F. Fabozzi. 2018. "Order from Chaos: How Data Science Is Revolutionizing Investment Practice." *The Journal of Portfolio Management* 45 (1): 1–4.

Simonian, J., and C. Wu. 2019. "Minsky vs. Machine: New Foundations for Quant-Macro Investing." *The Journal of Financial Data Science* 1 (2): 94–110.

Turkington, D., and A. Yazdani. 2020. "The Equity Differential Factor in Currency Markets." *Financial Analysts Journal* 76 (1): 70–81.

Disclaimer
The material presented is for informational purposes only. They reflect the current views of the author, are subject to change based on market and other conditions and factors, and do not necessarily represent the official views of State Street Corporation and its affiliations.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Minsky vs. Machine: *New Foundations for Quant-Macro Investing*

Joseph Simonian and Chenwei Wu
The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/2/94

ABSTRACT: *Systematic macro investors' use of the regime-switching models that have been developed in academia over the last several decades is infrequent at best and, when used, generally tangential to their core investment process. The roots of this less-than-enthusiastic uptake can be found in two familiar sources: models that possess an overly complex formal structure and poor predictive ability. As a remedy to the current state of affairs, the authors present a new foundation for regime-based investing, one based on spectral clustering, a graph theoretic approach to classifying data. Drawing inspiration from the work of Hyman Minsky and John Geanakoplos, the authors present a macro framework that uses measures of growth, inflation, and leverage to define regimes and drive portfolio decisions. To the latter end, the authors show how the framework can be used to build portfolios using information about regimes as defined, to outperform a no-information equal-weight portfolio both out-of-sample and in bootstrapped and cross-validated simulations. The authors thus show that spectral clustering can provide both an elegant mathematical description of the leverage cycle and a robust foundation for quant-macro investing.*

Order from Chaos: *How Data Science Is Revolutionizing Investment Practice*

Joseph Simonian, Marcos López de Prado, and Frank J. Fabozzi
The Journal of Portfolio Management
https://jpm.pm-research.com/content/45/1/1

ABSTRACT: *Data are fundamental inputs to any applied scientific endeavor. Finance is no different. Yet for almost its entire existence as an organized field of inquiry, much of finance has relied almost exclusively on relatively primitive and rigid forms of data analysis to drive both investment theories and real-world portfolio management decisions. Over the years, it has become evident that the complexity of today's markets, with a seemingly infinite amount of data being generated at rapid speeds, cannot be handled using the blunt mathematical arrows being deployed from the analytical quivers of most investment professionals. In response, the investment industry has begun to recognize the value and importance of the various methodological approaches that constitute what has come to be known as data science. Although the basic ideas that undergird much of data science can be traced at least back to Alan Turing (1947, 1948, 1951), it is only in recent decades, with rapid advances in computational power, that it has become possible to actualize many of the ideas of Turing and other computing pioneers.*

What Is an Index?

Andrew W. Lo
The Journal of Portfolio Management
https://jpm.pm-research.com/content/42/2/21

ABSTRACT: *Technological advances in telecommunications, securities exchanges, and algorithmic trading have facilitated a host of new investment products that resemble theme-based passive indexes, but depart from traditional market-cap-weighted portfolios. In this article, the author proposes broadening the definition of an index using a functional perspective. Any portfolio strategy that satisfies three properties should be considered an index: (1) it is completely transparent, (2) it is investable, and (3) it is systematic (i.e., it is entirely rules-based and involves no judgment or unique investment skill). Portfolios satisfying these properties that are not market-cap-weighted are given a new name: dynamic indices. This functional definition widens the universe of possibilities and, most importantly, decouples risk management from alpha generation. Passive strategies can and should be actively risk managed, and the author provides a simple example of how this can be achieved. Dynamic indices also create new challenges; the most significant of these is backtest bias, and the author concludes with a proposal for managing this risk.*

The Cross Section of Commodity Returns: *A Nonparametric Approach*

CLEMENS STRUCK AND ENOCH CHENG

CLEMENS STRUCK
is an assistant professor in the School of Economics at University College Dublin in Dublin, Ireland, and head of machine learning at PicardAngstAG in Switzerland.
clemens.struck@ucd.ie; clemens.struck@picardangst.com

ENOCH CHENG
is an assistant professor in the Department of Economics at the University of Colorado in Denver, CO.
enoch.cheng@ucdenver.edu

KEY FINDINGS

- Standard approaches to asset pricing make strong parametric assumptions that undermine their return-predicting ability.
- The authors employ tree-based methods to overcome these limitations and estimate the predictability of returns in commodities futures markets.
- Out of sample, they find that up to 3.74% of 1-month returns are predictable—more than a 10-fold increase from standard approaches.

ABSTRACT: *To what extent are financial market returns predictable? Standard approaches to asset pricing make strong parametric assumptions that undermine their return-predicting ability. The authors employ tree-based methods to overcome these limitations and attempt to approximate an upper bound for the predictability of returns in commodities futures markets. Out of sample, they find that up to 3.74% of 1-month returns are predictable—more than a 10-fold increase from standard approaches. The findings hint at the importance multiway interactions and nonlinearities acquire in the data; they imply that new factors should be tested on their ability to add explanatory power to an ensemble of existing factors.*

TOPICS: ***Futures and forward contracts, commodities****

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

The predictability of financial market returns is extensively discussed in the literature. Yet, the discussion predominantly centers around individual return factors and linear models à la Fama and French (1992). As a result of this focus, many studies make three strong parametric assumptions that undermine their return-predicting ability: (1) A simple functional form describes the statistical associations between prospective returns and predictors, (2) there are no interactions within the set predictors, and (3) the set usually includes less than a handful of predictors.

In this article, we compare the return-predicting ability of standard linear approaches to tree-based methods. The latter allow us to replace these ad hoc assumptions with data-driven counterparts.[1] More concretely, we examine the relation between 1-month returns and 59 predictors on a set of

[1] A discussion of this class of models, which is nonparametric and able to deal with many covariates, interactions, and nonlinearities, has been given by López de Prado (2018), Athey (2019), and Athey and Imbens (2019).

28 commodities futures.[2] Our fixed set of commodities futures is close to the efficient market ideal of information transparency, high liquidity, and low trading costs. In contrast to equity markets, it does not embody a survivorship bias problem and is thus perhaps better suited to address the aforementioned question.[3]

Out of sample, we document three main findings, which are validated by our robustness tests. First, up to 3.74% of 1-month returns are predictable using tree-based methods. Although still small, this number reflects a more than 10-fold increase from the standard linear approaches. Second, moving from a simple linear risk factor approach to tree-based methods raises the in-sample (out-of-sample) predictability from 0.58% to 56.21% (from 0.33% to 3.74%). Third, tree-based portfolios systematically outperform their linear counterparts not only in terms of absolute returns but also in risk-adjusted terms.

The appendices show that these findings are qualitatively robust to (A) altering the rebalancing frequency of the portfolio from 1 month to 2 weeks, (B) changing the forecasting horizon from 1 month to 2 weeks, (C) modifying the number of long and short positions in the portfolio, (D) employing a different pseudo random numbers generator, (E) adopting an alternative futures rolling procedure, (F) changing the cutoff between in-sample and out-of-sample period, (G) rebalancing the portfolio on the last business day of the month, and (H) normalizing the predictors on a zero-to-one scale.

Based on our main quantitative findings, we make two distinct points in a long debate dating back to Fama (1970) and the efficient market hypothesis. The latter posits that prospective returns cannot be predicted using factors that are constructed from past information unless these factors represent a risk. So far, the literature has separated out many individual factors that can account for a fraction of return variation; see, for example, Fama and French (1993, 2016); Longstaff (1995); Asness, Moskowitz, and Pedersen (2013); Novy-Marx (2013); and Farhi and Gabaix (2016)—to mention just a few popular ones. A key methodological problem is that most papers study only a small number of additive (linear) and independent (without interactions) factors in isolation. Our study contributes to this debate by showing that, out of sample, 10 times more return variation can be predicted than standard linear methods suggest in commodity futures markets. This finding is difficult to reconcile with an efficient markets perspective.[4]

Our second point is related. The fact that our results change quantitatively quite dramatically when moving from a simple linear risk factor approach to tree-based methods hints at a different structure in the data. Previous studies based on linear methods may thus offer qualitatively incorrect interpretations of the data. Consider, for example, the case of a single predictor that some study finds to be important in the sense that it is statistically significant when studied in isolation (or as an add-on to an oversimplified linear factor model). A possible qualitative conclusion could then be that this factor reflects a risk or a behavioral bias. However, this same factor may be unimportant (statistically insignificant) when studied among other predictors, allowing for multiway interactions and nonlinearities. The interpretation then would be that the first study discovered a false positive and that there is neither a risk nor a behavioral bias.

Our article contributes to a literature that has questioned the validity of many factor discoveries and stressed the possibility that some return anomalies are simply statistical artifacts. So far, many studies have

[2] We focus exclusively on commodities futures here, the key reason being that we want the largest possible selection of futures while using a fixed set of predictors that should be comparable across contracts. For example, an upward-sloping futures curve—one of the predictors we employ—may have a very different meaning in volatility futures markets than in commodities futures markets. Thus, we must focus on a single market here.

[3] In equity markets, researchers have to deal with sample selection problems that arise from new firms entering and old firms exiting the market on a rolling basis; see, for example, Elton, Gruber, and Blake (1996). It is difficult to address this issue in our context because the new firms typically have different characteristics compared to the old firms. A model that includes only the firms that survive would suffer from bias. At the same time, a model that is estimated (in sample) using old firms that drop out of the market and is then applied (out of sample) to new firms that have recently entered also suffers from bias.

[4] Owing to the statistically intuitive but difficult to interpret nature of tree-based methods, we cannot conclude whether our finding reflects a market risk or an inefficiency. The first explanation views return anomalies as a risk compensation and stipulates that asset pricing is rational and can be reconciled with a multifactor version of the capital asset pricing model (CAPM) or arbitrage pricing theory (APT). The second explanation views return anomalies as a mispricing. Explanations that follow this path include, among many others, De Bondt and Thaler (1990), De Long et al. (1990), MacKinlay (1995), Shleifer and Vishny (1997), Grinblatt and Han (2005), DellaVigna and Pollet (2009), and Hirshleifer (2015).

put the emphasis on sample selection bias and multiple testing problems; see, for example, Lo and MacKinlay (1990); Black (1993); Kothari, Shanken, and Sloan (1995); Jorion and Goetzmann (1999); Schwert (2003); Bailey and López de Prado (2014); Mclean and Pontiff (2016); Harvey, Liu, and Zhu (2016); and Harvey and Liu (2019). By contrast, we stress that false discoveries may also come from unprincipled model selection.

Our findings suggest an alternative procedure for testing new factors: Rather than testing whether a factor adds explanatory power in isolation, we should ask whether it contributes explanatory power when it is added to an ensemble of existing factors, taking into account multiway interactions and nonlinearities. Our point is thus similar to that of Feng, Giglio, and Xiu (2019), who proposed a linear two-step model selection procedure to evaluate new factors in a high-dimensional setting. Because we find multiway interactions and nonlinearities to be key in explaining the data, our results underline that a linear regularization procedure may not suffice.[5]

Our findings resonate well with recent tree-based approaches that study return factors in equity markets. Gu, Kelly, and Xiu (2020) compared various regression techniques and determined that a small subset of momentum-, liquidity-, and volatility-related factors are important in predicting asset returns. Their work focused on US equities examining 94 separate factors, with additional dummy variables including 74 different industry sectors, and interaction terms. Similar to Moritz and Zimmermann (2016), they found that nonlinearities play an important role in predicting returns. Both these papers demonstrate large gains to prediction when using tree-based forecasts over standard regression techniques.

What are the limitations of our results? Despite allowing for a sophisticated structure of the data, we are only able to match about 50% of the in-sample return variation. This may be due to two reasons. First, although our set of factors is larger than in previous studies, there is still room to extend it further—for example, to include interactions with macro factors such as exchange rate movements. Second, the two tree-based methods we employ do not allow for the most advanced structure possible. As we explain in the following, rather than finding globally optimal trees, we use a computationally more efficient procedure called the *greedy* algorithm in combination with a simple majority vote.

Both these simplifications indicate that there is further potential for return predictability. Yet, we still argue that our results approximate an upper bound. Here is why: Moving from standard linear techniques to tree-based methods increases the in-sample return predictability from 0.58% to 56.21%. Yet, in terms of out-of-sample predictability, this increase translates only into moving from 0.33% to 3.74%. Thus, increasing in-sample predictability has a limited effect on increasing out-of-sample predictability. We therefore believe that moving from an in-sample predictability of 56.21% to 100% may not increase the out-of-sample predictability much further.

Although our results do suggest that return predictability has been significantly underestimated by previous linear risk factor studies, they also show that over 95% of out-of-sample return variation cannot be predicted when taking into account a more sophisticated structure of the data via tree-based methods. In that sense, our findings are consistent with an efficient markets perspective that posits the unpredictability of returns using past information.

The article proceeds as follows. The second section explains the methodology and its implementation. The third section explains several setup choices we make and discusses the empirical findings. The fourth section interprets the main findings, and the fifth section concludes.

METHODOLOGY AND IMPLEMENTATION

We can formalize the problem of predicting returns as the response expected ahead of time t, $t + 1$, of commodity i, $E_t(y_{i,t+1})$, being a function, g, of a matrix of predictors $\mathbf{X}_{i,t} = (x_{i,t,1}, x_{i,t,2}, \ldots, x_{i,t,N})$. Formally,

$$E_t(y_{i,t+1}) = g(x_{i,t,1}, x_{i,t,2}, \ldots, x_{i,t,N}) = g(\mathbf{X}_{i,t}) \tag{1}$$

As we study the presence of systematic factors with consistent predictive power, we convert the problem of predicting returns into a classification problem of predicting positive and negative returns. This procedure is called *discretization* and lowers the impact of extreme returns, thereby reducing noise; see, for example, Fayyad and Irani (1993) and Garcia et al. (2013). We categorize the ahead return of asset i, $r_{i,\,t+1}$, into either one or zero,

[5] Feng, Giglio, and Xiu (2019) were mainly concerned with variable selection in a high-dimensional setting that included 150 factors. Their two-step LASSO procedure is linear and abstracts from interactions that we find to be crucial.

depending on whether it is positive or negative. More formally, the response variable, $y_{i,t+1}$, is coded as

$$y_{i,t+1} = \begin{cases} 1, & \text{if } r_{i,t+1} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

We now explain a variety of approaches (ranging from log-linear Logit and LASSO regressions to classification trees involving bagging and boosting) to model the function *g*. Later, we compare the empirical predictions of these different approaches.

Logit

As a baseline reference for standard linear methods, we use Logit. In the Logit regression, we model the probability of a response $y_{i,t+1}$ to our predictors $X_{i,t}$ through a log-linear function. Specifically, the probability of a positive forward return is assumed to be

$$\begin{aligned} &ln\left(\frac{Pr(y_{i,t+1}=1|\mathbf{X}_{i,t},\beta)}{1-Pr(y_{i,t+1}=1|\mathbf{X}_{i,t},\beta)}\right) \\ &= \beta_1 x_{i,t,1} + \beta_2 x_{i,t,2} + \ldots + \beta_N x_{i,t,N} \end{aligned} \tag{3}$$

where $E_t(y_{i,t+1}) = Pr(y_{i,t+1} = 1|X_{i,t}, \beta)$ and β is the $(N \times 1)$ parameter vector. Maximum likelihood is used to find the parameters that maximize the probability of observing the given data. Because $y_{i,t+1}$ is a binomial random variable, the probability density of $y_{i,t+1}$ is

$$f(y_{i,t+1} \mid \mathbf{X}_{i,t},\beta) = p_{i,t}^{y_{i,t+1}}(1-p_{i,t})^{1-y_{i,t+1}} \tag{4}$$

From the probability density function, we can derive the likelihood of different parameters given our observations. Oftentimes, this likelihood function is transformed into log-form to simplify calculation. Given the Logit probabilities

$$p_{i,t} = Pr(y_{i,t+1} = 1 \mid \mathbf{X}_{i,t},\beta) = \frac{\exp(\mathbf{X}'_{i,t}\beta)}{1+\exp(\mathbf{X}'_{i,t}\beta)} \tag{5}$$

whereby the β is chosen to maximize the log-likelihood function

$$\max_\beta \quad \mathcal{L} = \sum_{i=1}^{I}\sum_{t=1}^{T} y_{i,t+1}\ln(p_{i,t}) + (1-y_{i,t+1})\ln(1-p_{i,t}) \tag{6}$$

Later, in the empirical section, we back out the probabilities to estimate how well they predict the actual return cross section out of sample.

LASSO

A key problem of the Logit approach is that it requires us to select the predictors. For simplification purposes, in the empirical section, we estimate Logit using the entire set of predictors. In this case, estimation with many predictors may, however, suffer from multicolinearity problems. To bring a more systematic approach to variable selection, we employ an add-on to the Logit regression called LASSO. The LASSO is a modified log-linear regression that explicitly includes variable selection as an objective. It involves adding another term called the *shrinkage penalty* to the Logit's log-likelihood function.[6] Effectively, this shrinkage term sets βs of predictors that do not add information to zero and thereby acts as a variable selection tool. Formally, the new log-likelihood function is given by

$$\begin{aligned} \max_\beta \quad \mathcal{L} = &\sum_{i=1}^{I}\sum_{t=1}^{T} y_{i,t+1}\ln(p_{i,t}) \\ &+ (1-y_{i,t+1})\ln(1-p_{i,t}) - \lambda\sum_{n=1}^{N}|\beta_n| \end{aligned} \tag{7}$$

where λ is a scaling parameter that modifies the importance of the shrinkage penalty. We conduct a grid search and choose λ between 0 and 1 to maximize the log-likelihood.

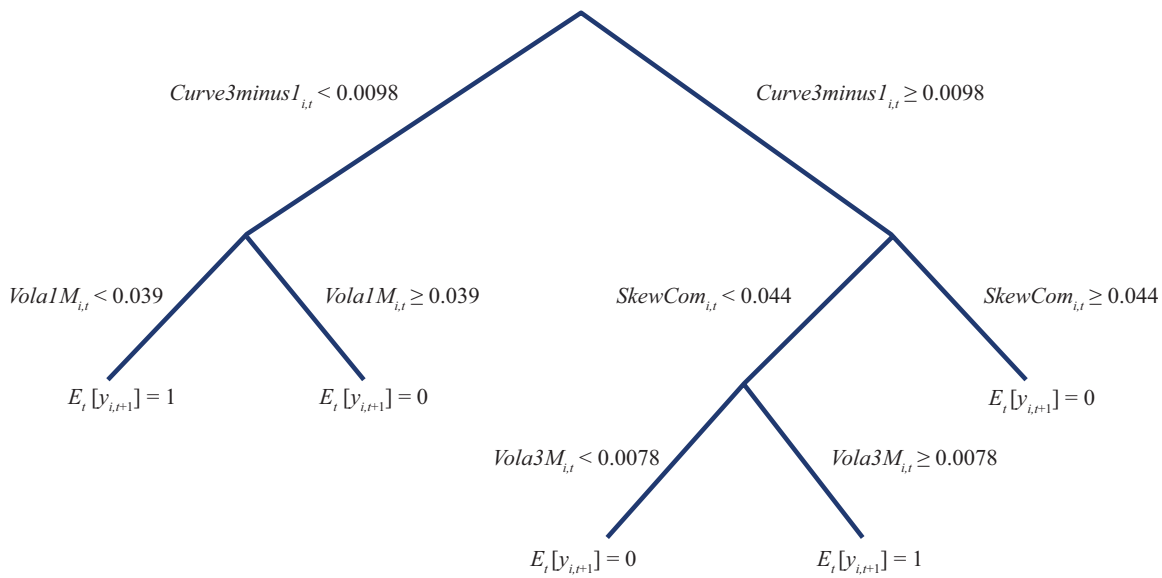
Ensemble Methods

A key problem of the previous two approaches is that they abstract from nonlinearities and interactions among the predictors. Regression trees are now a popular method in many fields to address these issues. We employ two variants of regression trees in this article. We begin by describing the construction of a sample tree and then proceed to explain the two variants: boosted and bagged trees.

Exhibit 1 portrays an example of a single decision tree drawn from our data. Because it is difficult to grow

[6] Technically, the MATLAB implementation minimizes a deviance function instead of maximizing the log likelihood. These two optimization problems have been shown to be equivalent.

EXHIBIT 1
A Decision Tree



these trees globally optimal, we employ a simple procedure called the greedy algorithm. Beginning from the top, each node j is associated with region R_j, which can be further split into two new regions, R'_j and R''_j. At the beginning of the procedure, we randomly select a subset of predictors at each node j. From this subset, a predictor n and its partitioning s is then chosen to minimize the fraction of $y_{i,t+1}$ misclassified by $E_t(y_{i,t+1})$. For every observation that may fall into a region R'_j or R''_j, the predicted response $E_t(y_{i,t+1})$ is determined by a majority vote in that region. For example, if there are more ones than zeros in region R'_j, then $E_t(y_{i,t+1}) = 1$, $\forall y_{i,t+1} \in R'_j$. More formally, the minimization procedure is given by

$$\min_{n,s} \quad \mathcal{L}_j = \frac{1}{N_{R'_j(n,s)}} \sum_{y_{i,t+1} \in R'_j(n,s)} \mathbb{1}(y_{i,t+1} \neq E_t(y_{i,t+1})) + \frac{1}{N_{R''_j(n,s)}} \sum_{y_{i,t+1} \in R''_j(n,s)} \mathbb{1}(y_{i,t+1} \neq E_t(y_{i,t+1})) \quad (8)$$

where $N_{R'_j(n,s)}$ is the total number of $y_{i,t+1}$ that fall into region $R'_j(n,s)$; R' is the region in which $x_{i,t,n} > s$, and R'' is the region where $x_{i,t,n} < s$. How do we stop growing trees? A typical stopping criterion is the point at which the resulting region contains less than a certain number of observations or when the last region contains only points of one class.

In Exhibit 1, the tree is constructed from four predictor variables: the third month versus front month futures term structure (*Curve3minus1*), the past 1-month volatility (*Vola1M*), the past 3 months' return volatility (*Vola3M*), and the past 260 trading days' return skew (*SkewCom*). There are two branches from the initial node, decided by partitioning *Curve3minus1*. On the left branch, *Vola1M* is partitioned, and on the right branch, *skewCom* is partitioned. There is one final partitioning of *Vola3M* on one of the branches of *skewCom*.

What is $E_t(y_{i,t+1})$ in the final regions? For example, if $Curve3minus1 < 0.0098$ and $Vola1M < 0.039$, we count all $y_{i,t+1}$ that fall into this region; if the majority of the $y_{i,t+1}$ take a value of unity, then $E_t(y_{i,t+1}) = 1$. If the majority of the $y_{i,t+1}$ in that region take a value of zero, then $E_t(y_{i,t+1}) = 0$. In our example, the function $g(X_{i,t})$ of Equation 1 is given by

$$\begin{aligned} E_t(y_{i,t+1}) = {} & \beta_1 \times \mathbb{1}(Curve3minus1 < 0.0098)\mathbb{1}(Vola1M < 0.039) \\ & + \beta_2 \times \mathbb{1}(Curve3minus1 < 0.0098)\mathbb{1}(Vola1M \geq 0.039) \\ & + \beta_3 \times \mathbb{1}(Curve3minus1 \geq 0.0098)\mathbb{1}(SkewCom < 0.044) \\ & \quad 1(Vola3M < 0.0078) \\ & + \beta_4 \times \mathbb{1}(Curve3minus1 \geq 0.0098)\mathbb{1}(SkewCom < 0.044) \\ & \quad 1(Vola3M \geq 0.0078) \\ & + \beta_5 \times \mathbb{1}(Curve3minus1 \geq 0.0098)\mathbb{1}(SkewCom \geq 0.044) \end{aligned}$$

Bagged trees. A main downside of classification trees is high variability. In other words, repeatedly growing an individual tree from the same dataset may result in different trees each time. In addition to variability, data fragmentation occurs. With each split, the nodes have successively fewer observations. Each path, on average, contains only a fraction of the predictor variables, implying that predictions are made by subsets of the predictors. Thus, we may accidentally cut out valuable predictors when growing a tree.

Several modifications have been developed to deal with these issues. A popular modification is called *ensemble* methods. *Bagging* is one such modification. It involves using a bootstrap to create different samples from the original data and then growing an individual tree on each bootstrapped sample (Breiman 1996; Dietterich 2000), thereby creating a *forest* (Breiman 2001; Criminisi, Shotton, and Konukoglu 2012). The predictions of the individual trees are then combined, again, by a simple majority vote. More formally, the combined prediction of B trees is

$$E_t(y_{i,t+1}) = \begin{cases} 1, & \text{if } \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(E_{t,b}(y_{i,t+1}) = 1) > 0.5. \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where $\mathbb{1}(E_{t,b}(y_{i,t+1}) = 1)$ is a binary indicator that takes the value of unity if the prediction of tree b, $E_{t,b}(y_{i,t+1})$, is unity and zero otherwise.

One problem with forests is that they tend to be homogeneous with regard to the order in which the predictor variables $x_{i,t,n}$ are selected. As mentioned previously, we address this issue by randomly choosing subsamples of the set of all predictors. In our case, we choose subsets that contain the square root of the number of predictor variables. The predictions made by bagged trees are more robust against overfitting because of the differences between bootstrapped samples.

Boosted trees. Boosting is an add-on to bagging in which the bootstrapped subsamples of data are not chosen at random. Instead, bagging and growing trees takes place sequentially. After a tree is grown, it is tested on the entire sample of data. In the next bootstrapped sample, the previously incorrectly classified observations are then overrepresented, with the effect that the next tree is chosen to minimize the prediction errors of all previous trees; see, for example, Schapire and Singer (1999). As each tree is grown, the forest incrementally improves prediction accuracy for the forest. The variant of boosting that we employ is called gentle adaptive boosting.

Instead of using the predicted class for boosted and bagged trees, we back out the probabilities of being in one class or the other. This procedure allows us later to sort our commodities set in each period.

EMPIRICAL ANALYSIS

Data

We focus on a fixed universe that consists of all liquid commodities futures that are tradable throughout our sample period of January 1987 to October 2019. The data are drawn from Bloomberg.[7] By choosing a fixed set of futures, we avoid dealing with problems that arise from survivorship bias—a pervasive phenomenon in equity markets. The liquidity of the futures curve expands throughout the sample period. To maintain comparability across time, we focus only on front-, second-, and third-month contracts. We employ the latter two contracts only to construct a signal. Trading is entirely based on front-month contracts.

Commodities. The commodities are Soybean Oil CBOT (BO); Corn CBOT (C); Sugar ICE (QW); Cocoa ICE (CC); WTI Crude Oil Nymex (CL); Brent Crude Oil ICE (CO); Cotton #2 ICE (CT); Feeder Cattle CME (FC); Gold Comex (GC); Copper Comex (HG); Heating Oil Nymex (HO); Coffee 'C' ICE (KC); Kansas Wheat CBOT (KW); Aluminum LME (LA); Live Cattle CME (LC); Lean Hogs CME (LH); Lead LME (LL); Nickel LME (LN); Zinc LME (LX); Natural Gas Nymex (NG); Platinum Nymex (PL); Gasoil Nymex (QS); Soybeans CBOT (S); Sugar #11 ICE (SB); Silver Comex (SI); Soybean Meal CBOT (SM); Wheat CBOT (W); and RBOB Gasoline Nymex (XB).

To construct continuous excess return series of daily closing prices for each commodity, we employ a standard futures rolling procedure. If the last trade date for a contract is in the month before the contract month, we roll 10 days before the last trade date. For some

[7] Because the computations are quite demanding, we use MATLAB Parallel Server in combination with an Amazon Web Services cluster with one stack of 96 virtual CPUs. This allows us to complete the computations within a single day.

commodities, first notice is before the last trade date, and there is not much trading activity after first notice. Thus, we roll 10 days in advance to ensure liquidity. Consider, for example, a contract that points to the Brent Crude September 2019 contract. The last trade date for the September contract is July 31, 2019. Because the last trade date is before the contract month, the contract would roll 10 days before the last trade date, July 31, 2019. If the last trade date for a contract is in the month during or after the contract month, we roll 10 days before the first day of the contract month. Take, for example, a contract that points to the Corn December 2019 contract. The last trade date for the December contract is December 13, 2019. Because the last trade date is during the corresponding contract month, the contract would roll 10 days before the first day of the contract month, December 1, 2019. If the first day of the contract month is a nonbusiness day, it will roll 10 days before the next business day.

Online Appendix E shows the robustness of our results employing an alternative futures rolling procedure.

Predictors

A characteristic of commodity futures is that prices are nearly a pure reflection of economic fundamentals—spot prices and storage costs as well as current future demand and supply. A key strength of our approach is selection of predictors that are constructed from objectively measurable historical market prices. Our measures are motivated by previous studies highlighting factors such as volatility, momentum, value, size, liquidity, and option implied tail risks; see, for example, Sharpe (1964); Fama and French (1992); Longstaff (1995); Asness, Moskowitz, and Pedersen (2013); and Farhi and Gabaix (2016), among others. Fuertes, Miffre, and Fernandez-Perez (2015) studied risk factors in commodity futures markets employing several of these factors while adding past return skewness and kurtosis.

Momentum. Our first predictors, $x_{i,t,1-4}$, are variants of the same time-series momentum signal—a commodity's own past return (see, e.g., Moskowitz, Ooi, and Pedersen 2012). Formally,

$$x_{i,t,1-4} = \frac{P_{i,t}}{P_{i,t-q}} - 1 \qquad \forall q \in \{10, 20, 60, 180\} \tag{10}$$

where q denotes the number of workdays and determines the length of the look-back window, and $P_{i,t}$ is the level of the excess return series of commodity i at time t. Our second set of predictors also consists of momentum measures, this time constructed from moving averages; see, for example, Brock, Lakonishok, and Lebaron (1992). More specifically, they measure the convergence of short- and long-term moving averages. Formally,

$$x_{i,t,5-8} = \frac{\left(\frac{1}{Q}\sum_{q=0:Q} P_{i,t-q} - \frac{1}{Z}\sum_{z=0:Z} P_{i,t-z}\right) \Big/ \left[\frac{1}{Z}\sum_{z=0:Z} P_{i,t-z}\right]}{\left(\frac{1}{Q}\sum_{q=10:(Q+10)} P_{i,t-q} - \frac{1}{Z}\sum_{z=10:(Z+10)} P_{i,t-z}\right) \Big/ \left[\frac{1}{Z}\sum_{z=10:(Z+10)} P_{i,t-z}\right]} - 1, \tag{11}$$

where the look-back window of the long-term moving average, Z, is either 260 or 360 weekdays, and the look-back window of the short-term moving average, Q, is either 30 or 60 weekdays.

Volatility. Our third set of predictors measures past daily volatility and is thus similar to the risk measure hypothesized by Sharpe (1964). Formally,

$$x_{i,t,9-12} = \sqrt{\frac{1}{Q-1}\sum_{q=0:Q}\left(\frac{P_{i,t-q} - P_{i,t-1-q}}{P_{i,t-1-q}} - \mu_Q\right)^2} \qquad \forall Q \in \{10, 20, 60, 180\} \tag{12}$$

where μ_Q denotes the average daily return; the look-back window, Q, is given by the number of weekdays.

Value. To capture value, we use three different indicators. First, we use two measures of the steepness of the futures curve. Intuitively, the futures curve is typically in contango (upward sloping) to account for storage costs. A steeper contango could indicate a temporary excess supply of the commodity, corresponding to a temporarily lower price. Vice versa, a futures curve that is in backwardation (downward sloping) could indicate a temporary scarcity of the commodity, corresponding to a temporarily higher price. Formally,

$$x_{i,t,13-14} = \frac{F_{q,t} - F_{1,t}}{F_{1,t}} \qquad \forall q \in \{2, 3\}, \tag{13}$$

where q denotes the contract month (e.g., $F_{3,t}$ would be the third-month futures price at time t). Schwerhoff and Stuermer (2019) showed that long-term prices for commodities are rather stable. Thus, we also employ a value measure that aims to capture the current price relative to a long-term historical average. Formally,

$$x_{i,t,15} = \frac{F_{1,t}}{\frac{1}{4 \times 260} \sum_{q=0:4\times 260} F_{1,t-q}} \tag{14}$$

The look-back window is admittedly a bit ad hoc but is similar to the one chosen by Asness, Moskowitz, and Pedersen (2013).

Miscellaneous. The following measures do not fall into any of the previous categories. One such measure captures the risk reversal and was studied by Farhi and Gabaix (2016) in the context of currency markets. More specifically, it is the difference in the implied 30-day volatility of out-of-the-money 95% put and 105% call options. Intuitively, the differences in option implied volatility directly reflects a difference in the option prices, which may indicate that the 1-month return risk is skewed either to the right or left. Formally,

$$x_{i,t,16} = \mathrm{iVol}^{30\,\mathrm{days}}_{95,i,t} - \mathrm{iVol}^{30\,\mathrm{days}}_{105,i,t} \tag{15}$$

Following Fuertes, Miffre, and Fernandez-Perez (2015), we also employ skewness and the kurtosis of the daily returns of the past year. Formally,

$$x_{i,t,17} = \frac{\frac{1}{Q}\sum_{q=0:Q}\left(\frac{P_{i,t-q} - P_{i,t-1-q}}{P_{i,t-1-q}} - \mu_Q\right)^3}{\left(\sqrt{\frac{1}{Q}\sum_{q=0:Q}\left(\frac{P_{i,t-q} - P_{i,t-1-q}}{P_{i,t-1-q}} - \mu_Q\right)^2}\right)^3} \tag{16}$$

and

$$x_{i,t,18} = \frac{\frac{1}{Q}\sum_{q=0:Q}\left(\frac{P_{i,t-q} - P_{i,t-1-q}}{P_{i,t-1-q}} - \mu_Q\right)^4}{\left(\sqrt{\frac{1}{Q}\sum_{q=0:Q}\left(\frac{P_{i,t-q} - P_{i,t-1-q}}{P_{i,t-1-q}} - \mu_Q\right)^2}\right)^4} \tag{17}$$

where Q is set to 260 weekdays. The last measure in this category is the market size of each commodity relative to other commodities. We take the moving-average past trading volume relative to the trading volume of a basket of eight core commodities, $j \in$ {BO, C, CL, HO, LC, S, SM, W}. Formally,

$$x_{i,t,19} = \frac{1}{Z}\sum_{z=0:Z} \mathrm{relVol}_{i,t-z} \tag{18}$$

where

$$\mathrm{relVol}_{i,t} = \frac{\frac{1}{Q}\sum_{q=0:Q}\mathrm{Vol}_{i,t-q}}{\frac{1}{Q}\frac{1}{j}\sum_{q=0:Q}\sum_{j}\mathrm{Vol}_{j,t-q}}.$$

$\mathrm{Vol}_{i,t}$ is the trading volume in US dollars of commodity i at time t. We set Q and Z to 40 weekdays.

Dummies. Finally, we also employ a set of dummies. First, to account for the seasonality of some commodities, such as NG, we employ 12 dummies, one for each month. Second, to account for commodity-specific behavior, we employ 28 dummies, one for each commodity.

In total, we use a set of 59 predictors. Data scaling is a recommended preprocessing step when working with deep learning neural networks; see, for example, Bishop (1995). A legitimate concern is that some implementations of the regression techniques that we use may assign greater weight to predictors that take higher absolute values. In Online Appendix H, we follow the procedure of Struck and Cheng (2019) and estimate the empirical distribution function of each predictor using a kernel estimator and then use the inverse cumulative distribution function to normalize the variables on a zero-to-one scale. The quantitative results change only marginally and do not affect any of our conclusions.

Parameterization

We rely on standard implementations of the algorithms. Because standard bagged tree estimators are highly sensitive to the underlying random number stream,[8] we repeat the estimation procedure nearly

[8] Intuitively, with 59 predictors, there are billions of trees that can be constructed from the data; our forest contains a small subset of them.

100 times, starting the same random number stream at different points to ensure our results are robust. We average the resulting probabilities over all estimates. In addition, Online Appendix D shows the results if we estimate repeatedly from a different type of random number stream. For bagged trees, we set the number of trees close to 500. We stress that the number of drawn trees does not qualitatively change our results. For boosted trees, we set the learning rate to 0.5—a standard parameter between 0 and 1. For LASSO, we allow for a 10-fold cross-validation—standard in the literature.

To avoid multiple testing problems, we present the results using our initial cutoff for the in-sample and out-of-sample period: January 1987 to December 2013 and January 2014 to October 2019. Online Appendix F presents a robustness test with an alternative cutoff point (September 2016). In both cases, we choose a rather long in-sample period. There are two main reasons: First, as Rietz (1988), Barro (2006), Farhi and Gabaix (2016), and Struck and Cheng (2019) underlined, tail events are key for the performance of many factor strategies. The in-sample period should therefore include as many tail events as possible. Second, we want to show that our findings are robust in recent years before publication and that we do not describe some phenomenon that has already been arbitraged out.

Because long-run commodity prices are relatively stable over the entire sample period (possibly due to innovations that increase the supply in the light of increasing global demand for commodities; see, e.g., Acemoglu 2002), we construct market-neutral portfolios by sorting commodities according to their predicted return probabilities. We choose five equally weighted long and five equally weighted short positions from the top and bottom of this sorting. Online Appendix C shows the robustness of our results when using three long and three short positions instead. To benchmark different risk factor strategies, we also construct an equally weighted long-only strategy using the 28 commodities futures.

We employ a standard forecast horizon of 4 weeks; see, for example, Moskowitz, Ooi, and Pedersen (2012). Online Appendix B illustrates the robustness of our results when the forecast horizon is 2 weeks instead. Given the forecast horizon, we also choose to rebalance every 4 weeks. Online Appendix A illustrates the robustness of our results when rebalancing is every 2 weeks instead. Online Appendix G shows the robustness of our results when rebalancing of the portfolios takes place on the last business day of the month.[9]

Finally, given that we trade only liquid futures contracts at a low frequency, we abstract from transaction costs. These costs amount to less than 8 bps per annum and affect the returns of all strategies in the same way.

Main Findings

Our empirical analysis proceeds in four steps. First, we compare the ability of the four risk factor models (Logit, LASSO, bagged trees, boosted trees) to predict subsequently realized returns in sample and out of sample. The results are shown in Exhibit 2. Second, we compare the out-of-sample performance of the four risk factor models and the long-only strategy. The results are presented in Exhibits 3 and 4. Third, we focus on the individual importance that each model assigns to the predictors. A ranking is presented in Exhibit 5. Finally, we focus on the correlations of the tree-based portfolios with other asset classes and the standard linear risk factor approach. The results are presented in Exhibits 6 and 7.

More concretely, Exhibit 2 shows the pooled correlation (across commodities returns all employing 4-week intervals) between the ex ante probabilities at the beginning of each 4-week interval and the actual realized returns over the subsequent 4-week period. Rows i and ii are the correlations using all 28 commodities. Rows iii and iv show the correlations using only the top five and bottom five commodities (sorted by ex ante probabilities) of each interval. Rows v and vi show the correlations using only the top three and bottom three commodities (sorted by ex ante probabilities) of each interval. Intuitively, the squared correlation captures the percentage of returns that is predicted by the different methods that we employ.

There are five key takeaways from Exhibit 2. First, out of sample, the squared correlation is up to 3.74% in the top-three/bottom-three sample. Second, out of sample, using bagged trees results in a squared correlation that exceeds linear risk factor approaches (Logit, LASSO) by a factor of at least 10. Third, out-of-sample predictability increases in the extremes across

[9] Some asset managers use this rebalancing frequency to simplify their reporting, among other reasons.

EXHIBIT 2

Squared-Correlation between 4W Signals and 4W Returns (main result)

	Logit	LASSO	Bagged Trees	Boosted Trees
i. In Sample, All Commodities	0.42%	0.42%	50.26%	51.87%
ii. Out of Sample, All Commodities	0.00%	0.00%	0.99%	0.33%
iii. In Sample, Top 5/Bottom 5	0.61%	0.66%	55.41%	54.59%
iv. Out of Sample, Top 5/Bottom 5	0.12%	0.10%	2.02%	1.40%
v. In Sample, Top 3/Bottom 3	0.58%	0.57%	56.21%	56.23%
vi. Out of Sample, Top 3/Bottom 3	0.33%	0.25%	3.74%	1.44%

Notes: Four-week forecast horizon, rebalancing every four weeks. In-sample period: January 1987 to December 2013; out-of-sample period: January 2014 to October 2019. Standard futures rolling procedure. Pseudo random numbers generator: CombRecursive.

EXHIBIT 3

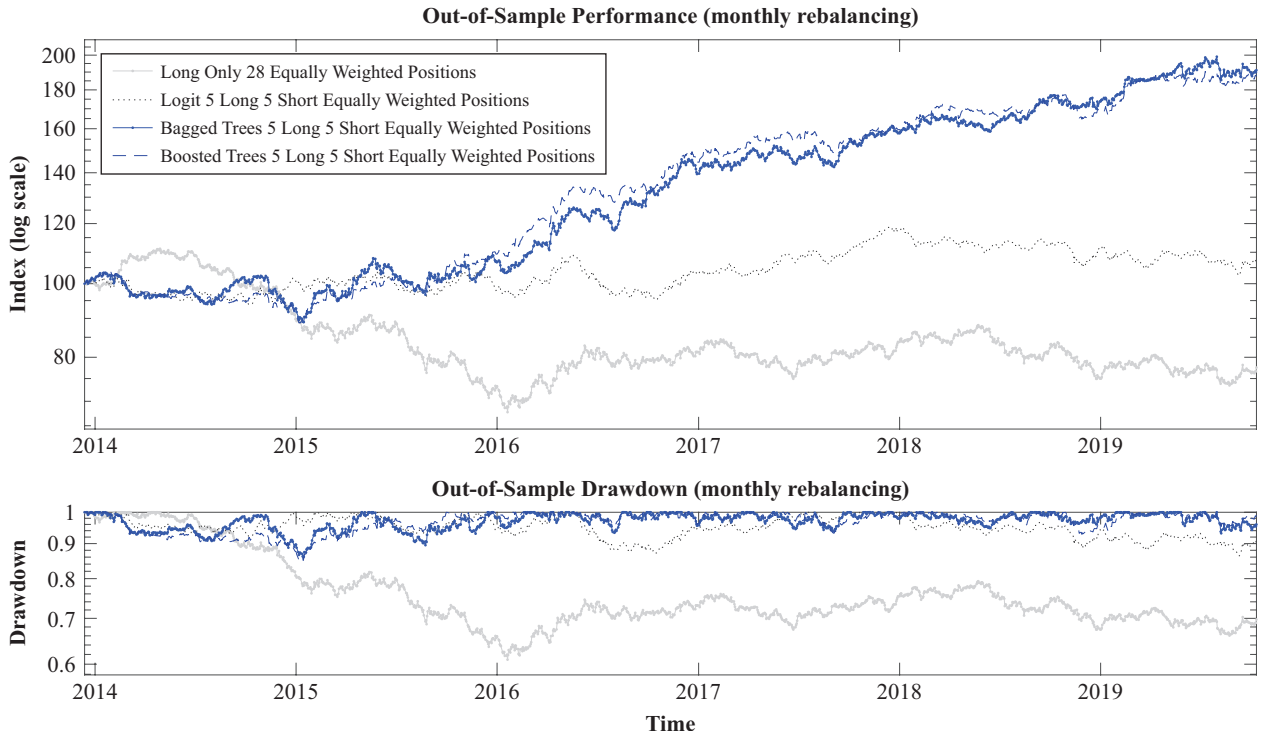
Historical Out-of-Sample Performance: Machine Learning vs. Linear Risk Factor Approaches vs. Long Only (main result)

	Sample	Last 3 Years	Last 2 Years	Last 1 Year	Year-to-Date
Boosted Trees: Sharpe Ratio	1.296	1.343	1.224	0.736	1.562
Return-to-|MaxDD|	0.75	1.644	1.443	0.936	2.007
Volatility	0.087	0.088	0.085	0.091	0.092
Return	0.112	0.118	0.104	0.067	0.144
Max Drawdown	–0.15	–0.072	–0.072	–0.072	–0.072
Bagged Trees: Sharpe Ratio	1.101	1.34	1.193	0.79	1.118
Return-to-|MaxDD|	0.836	1.916	1.708	1.33	1.86
Volatility	0.106	0.098	0.098	0.115	0.114
Return	0.116	0.131	0.117	0.091	0.127
Max Drawdown	–0.139	–0.068	–0.068	–0.068	–0.068
LASSO: Sharpe Ratio	0.174	0.443	–0.221	–0.489	–0.407
Return-to-|MaxDD|	0.115	0.274	–0.136	–0.349	–0.275
Volatility	0.092	0.086	0.085	0.099	0.094
Return	0.016	0.038	–0.019	–0.048	–0.038
Max Drawdown	–0.139	–0.139	–0.139	–0.139	–0.139
Logit: Sharpe Ratio	0.187	0.526	–0.168	–0.439	–0.461
Return-to-|MaxDD|	0.125	0.328	–0.104	–0.309	–0.312
Volatility	0.092	0.086	0.085	0.097	0.093
Return	0.017	0.045	–0.014	–0.042	–0.043
Max Drawdown	–0.137	–0.137	–0.137	–0.137	–0.137
Long-Only: Sharpe Ratio	–0.353	–0.032	–0.218	–0.609	0.487
Return-to-|MaxDD|	–0.097	–0.009	–0.063	–0.182	0.14
Volatility	0.107	0.096	0.098	0.102	0.097
Return	–0.038	–0.003	–0.021	–0.062	0.047
Max Drawdown	–0.391	–0.339	–0.339	–0.339	–0.339

Notes: Four-week forecast horizon; rebalancing every four weeks. Five long, five short, equally weighted. In-sample period: January 1987 to December 2013; out-of-sample period: January 2014 to October 2019. Standard futures rolling procedure. Pseudo random numbers generator: CombRecursive.

EXHIBIT 4

Historical Out-of-Sample Performance: Machine Learning vs. Linear Risk Factor Approaches vs. Long Only (main result)



Notes: Four-week forecast horizon; rebalancing every four weeks. Five long, five short, equally weighted. In-sample period: January 1987 to December 2013; out-of-sample period: January 2014 to October 2019. Standard futures rolling procedure. Pseudo random numbers generator: CombRecursive.

all methods. For example, the predictability of bagged trees in the top-three/bottom-three sample is higher (3.74%) than the predictability in the top-five/bottom-five sample (2.02%) and the predictability in the all-commodities sample (0.99%). Fourth, the tree-based methods that we employ achieve a much higher in-sample squared correlation than the linear risk factor methods. Finally, the high in-sample squared correlation of the tree-based methods does not translate into an equally high out-of-sample squared correlation.

Exhibit 3 describes the out-of-sample behavior of the different strategies. The exhibit shows two risk-adjusted return measures: the Sharpe ratio (computed from daily volatility and daily returns) and the return-to-drawdown ratio. It also shows the average annualized daily returns and volatility as well as the maximum drawdown during the corresponding period. The out-of-sample period is broken down into five subperiods: the entire out-of-sample period, last 3 years, last 2 years, last 1 year, and year-to-date. Exhibit 4 visualizes these results.[10]

There are two noteworthy results. First, both tree-based methods (bagged and boosted trees) clearly outperform linear risk factor methods (Logit and LASSO), and the long-only approach in terms of risk-adjusted returns (Sharpe ratio and return to maximum drawdown) for all horizons (entire out-of-sample period, last 3 years, last 2 years, last 1 year, year-to-date). This outperformance is sizable. Whereas the linear risk factor methods feature a Sharpe ratio between −0.461 and 0.526, the Sharpe ratio of the tree-based methods lies between 0.736 and 1.562. The Sharpe ratio of the long-only strategy lies between −0.609 and 0.487. Furthermore, the return-to-drawdown ratio of the linear risk factor methods varies between −0.349 and 0.274, whereas the same measure is in a much higher range for the tree-based

[10] We drop the LASSO performance from the exhibit because it closely resembles the performance of the Logit strategy.

EXHIBIT 5
Predictor Importance

	Logit/LASSO	Boosted Trees	Bagged Trees
1	Ret3M ($x_{i,t,3}$)	Curve3minus1	Comvalue
2	Comvalue ($x_{i,t,15}$)	Curve2minus1	KurtCom
3	KurtCom ($x_{i,t,18}$)	Ret2W	Vola3M
4	Ret1M ($x_{i,t,2}$)	Vola2W	Vola9M
5	SkewCom ($x_{i,t,17}$)	KurtCom	SlMovAvg60D260D
6	Vola3M ($x_{i,t,11}$)	Ret1M	VolumeCom
7	Vola1M ($x_{i,t,10}$)	SkewCom	Vola1M
8	Skewivol ($x_{i,t,16}$)	Vola1M	SlMovAvg60D360D
9	VolumeCom ($x_{i,t,19}$)	Vola3M	Vola2W
10	Vola9M ($x_{i,t,12}$)	Ret3M	Skewivol
11	Vola2W ($x_{i,t,9}$)	VolumeCom	SlMovAvg30D260D
12	Curve3minus1 ($x_{i,t,14}$)	Vola9M	SkewCom
13	SlMovAvg30D260D ($x_{i,t,5}$)	Ret9M	SlMovAvg30D360D
14	Ret2W ($x_{i,t,1}$)	SlMovAvg60D260D	Ret2W
15	SlMovAvg60D360D ($x_{i,t,8}$)	SlMovAvg30D260D	Curve3minus1
16	Ret9M ($x_{i,t,4}$)	Comvalue	Ret1M
17	SlMovAvg60D260D ($x_{i,t,6}$)	SlMovAvg60D360D	Ret9M
18	Curve2minus1 ($x_{i,t,13}$)	SlMovAvg30D360D	Ret3M
19	SlMovAvg30D360D ($x_{i,t,7}$)	Skewivol	Curve2minus1

Notes: Four-week forecast horizon. In-sample period: January 1987 to December 2013. Standard futures rolling procedure. Pseudo random numbers generator: CombRecursive.

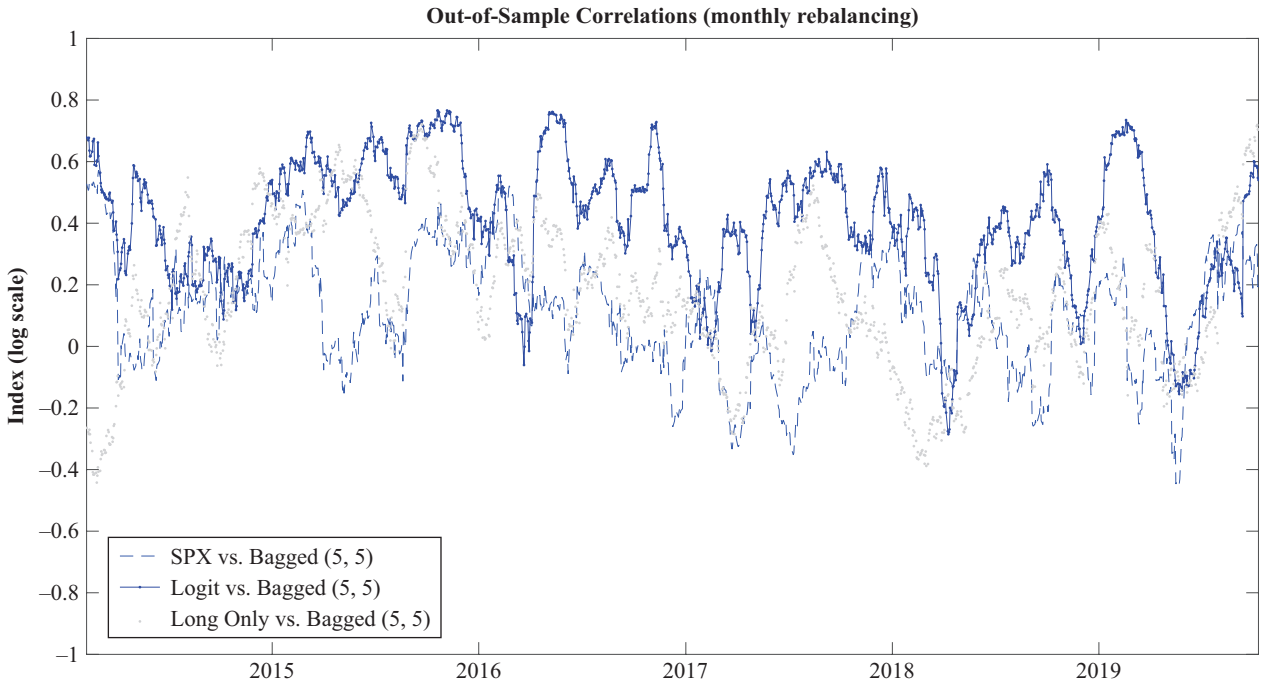
methods—between 0.75 and 2.007. The long-only strategy's return-to-drawdown ratio lies between −0.182 and 0.140. Moritz and Zimmermann (2016) and Gu, Kelly, and Xiu (2020) both documented large risk-adjusted gains over linear approaches when studying US equity markets. Although difficult to compare because of the different markets and time span, their quantitative findings are broadly coherent with ours.

Second, a closer examination of Exhibits 3 and 4 reveals that the outperformance in risk-adjusted terms is mainly driven by higher returns and less so by lower volatility or better drawdown behavior. Both tree-based methods outperform the linear risk factor approaches and long only by at least 8.0 percentage points per annum without compromising on drawdown behavior or volatility. In contrast, the volatility of the tree-based methods is only up to 2.1 percentage points higher than the volatility of the linear risk factor approaches (in the "year-to-date" column, bagged trees have a volatility of 11.4% whereas Logit has a volatility of only 9.3%). It can also be 0.8 percentage points lower (e.g., in the "last 1 year" column, the volatility of boosted trees is 9.1% whereas the volatility of LASSO is 9.9%). Similarly, the maximum drawdown of tree-based methods can be 1.3 percentage points worse than the maximum drawdown of linear risk factor methods (in the "sample" column, boosted trees is −15% whereas Logit is only −13.7%). However, the maximum drawdown can also be 7.1 percentage points better (in the "last 3 years" column, bagged trees is only −6.8% whereas LASSO has a drawdown of −13.9%).

When we examine the ranking of predictor importance, we note that there are few similarities across the different methods (Exhibit 5). Value ($x_{i,t,16}$), 1M and 3M momentum ($x_{i,t,2}$ and $x_{i,t,3}$), return skewness ($x_{i,t,17}$), and kurtosis, ($x_{i,t,18}$) are the top five predictors according to the linear risk factor approaches. Under boosted trees, four of these predictors are among the top 10 (1M and 3M momentum, return skewness, and kurtosis). Under bagged trees, two of them (value and return kurtosis) rank among the top five. Generally, however, the methods rank predictors differently in terms of importance. Here, our results differ from those of Gu, Kelly, and Xiu (2020), who found quite consistent rankings across all regression techniques in US equity markets.

EXHIBIT 6

Historical Out-of-Sample Correlations: Machine Learning vs. Linear Risk Factor Approaches vs. Long Only (main result)



Notes: Four-week forecast horizon; rebalancing every four weeks. Five long, five short, equally weighted. In-sample period: January 1987 to December 2013; out-of-sample period: January 2014 to October 2019. Standard futures rolling procedure. Pseudo random numbers generator: CombRecursive.

The dissimilarities we document in these rankings are further reflected in a low correlation among the strategies. Exhibits 6 and 7 show the pairwise 40-day rolling correlations for bagged and boosted trees, respectively. The correlations in Exhibit 6 between bagged trees and Logit vary between −0.2 and 0.8 with a long-term average near 0.35. The correlations in Exhibit 7 between boosted trees and Logit vary between −0.6 and 0.8 with an average near 0.2. Thus, in addition to the difference in performance between tree-based methods and linear risk factor approaches, these correlations in conjunction with the predictor importance ranking indicate that the actual structure of the data is quite different from the one suggested by simple linear risk factor approaches.

Moreover, the two exhibits also reveal that the tree-based correlations with a long-only commodity portfolio and the stock market (S&P 500) are low. Portfolios constructed based on tree-based methods may therefore offer additional diversification benefits for long-only stock portfolios that are exploited by neither linear risk factor strategies nor by long-only commodity strategies.

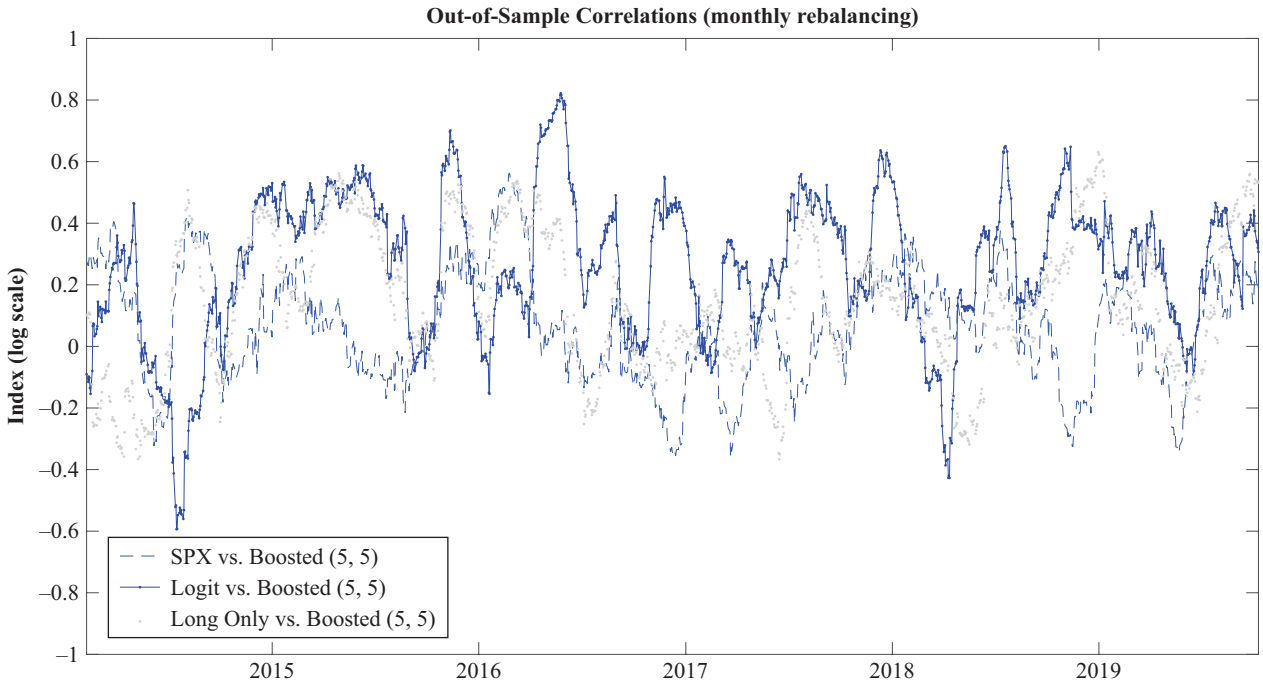
DISCUSSION

How do we interpret our findings in the light of previous studies? Based on our main quantitative findings, we make two distinct points in a long debate dating back to Fama (1970) and the efficient market hypothesis. First, markets may be less efficient than suggested by previous studies. Second, the usual approach of studying factors in isolation may lead to misinterpretations of the data, including the discovery of false positives.

Our first point builds on our main finding in studying a market that closely resembles the efficient market ideal of information transparency, high liquidity, and low transaction costs—commodities futures. Because we study a fixed set of commodities futures, we avoid dealing with the survivorship bias that affects

EXHIBIT 7

Historical Out-of-Sample Correlations: Machine Learning vs. Linear Risk Factor Approaches vs. Long Only (main result)



Notes: Four-week forecast horizon; rebalancing every four weeks. Five long, five short, equally weighted. In-sample period: January 1987 to December 2013; out-of-sample period: January 2014 to October 2019. Standard futures rolling procedure. Pseudo random numbers generator: CombRecursive.

equity market studies.[11] Furthermore, in contrast to equity market studies, we use only factors that are based on historical market prices and can therefore be precisely measured.[12]

So far, the literature has identified many individual factors that can explain a tiny fraction of the return variation. However, most papers study only a small number of additive (linear) and independent (without interactions) factors in isolation. What those factors represent economically is an ongoing debate. The efficient market hypothesis posits that prospective returns cannot be predicted using factors that are constructed from past information unless they reflect a risk. Our study contributes to this debate by showing that 10 times more return variation can be predicted than linear approaches suggest when using a large set of factors taking into account multifactor interactions and nonlinearities. Our findings pass a large number of robustness tests.

In light of the statistically intuitive but difficult to interpret methods that we use, we cannot conclude whether our finding reflects a market risk[13] or an inefficiency. Yet, it suggests that much more return variation is predictable than previously assumed, which is difficult to reconcile with an efficient markets perspective.

Our second point also builds on our main finding that return predictability is much higher than previous studies suggest. As alluded to earlier, the overwhelming majority of the literature has made statistically oversimplifying assumptions when studying return predictability. The intention of many of these papers has been to discover and study new predictors in isolation.

[11] Elimination of survivorship bias is crucial when quantifying the relation between factors and prospective returns; see, for example, Elton, Gruber, and Blake (1996).

[12] A common problem in equity market studies is that they employ factors that have been hypothesized in the past and remeasured again and again until they have explanatory power. This is a form of confirmation bias.

[13] A market risk would imply that our finding can be reconciled with an extension of the CAPM or APT and thus the efficient market hypothesis itself.

The oversimplified statistical frameworks that such papers adopt suits that purpose to the extent that their results are easy to interpret. However, in light of our study, it is clear that two additional sources of return predictability—multiway interactions among predictors and nonlinearities—are also important.

Previous studies that abstract from these two sources may thus offer qualitatively incorrect interpretations of the data. Consider, for example, the case of a single predictor that some study finds to be important in the sense that it is statistically significant when it is studied in isolation. A possible qualitative conclusion could then be that this factor reflects a risk or a behavioral bias. However, this same factor may be unimportant (statistically insignificant) when studied jointly with other predictors, allowing for multiway interactions and nonlinearities. The interpretation then would be that the first study discovered a false positive and that there is neither a risk nor a behavioral bias.

The fact that our results differ quantitatively when we move from a simple linear risk factor approach to tree-based methods hints at a different structure in the data. Thus, apart from being able to better quantify the predictability of returns, we argue that our empirical results warn of simplistic interpretations derived from studies that employ a few or even just a single factor. Our study indicates that claims about individual factors derived in isolation may not hold up well in practice.

What are the limitations of our results? Despite allowing for a sophisticated data structure, Exhibit 2 shows that we are only able to match about 50% of the in-sample return variation. This may be due to two reasons. First, although our set of factors is larger than in previous studies, there is still room to extend it further—for example, to include interactions with macro factors such as exchange rate movements. Second, the two tree-based methods we employ do not allow for the currently most advanced structure available. As we stressed earlier, rather than finding globally optimal trees, we use a computationally more efficient procedure called the greedy algorithm in combination with a simple majority vote.

Both of these simplifications indicate that there is further potential for return predictability. Yet, we still argue that our results approximate an upper bound. Our argument is simple: Moving from standard linear techniques to tree-based methods increases the in-sample return predictability from 0.58% to 56.21% (in the top-three/bottom-three sample). Yet, in terms of out-of-sample predictability, this increase translates only into moving from 0.33% to 3.74%. Thus, increasing in-sample predictability has a limited effect on increasing out-of-sample predictability. We therefore believe that moving from an in-sample predictability of 56.21% to 100% may not increase the out-of-sample predictability much further.

Although our results do suggest that return predictability has been significantly underestimated by previous linear risk factor studies, they also show that over 95% of out-of-sample return variation cannot be predicted when taking into account a more sophisticated data structure via tree-based methods. In that sense, our findings are consistent with an efficient markets perspective that posits the unpredictability of returns using past information.

CONCLUSIONS

To what extent are financial market returns predictable? Common linear approaches à la Fama and French (1992) impose three ad hoc assumptions on the data that undermine their return-predicting ability. In this article, we compare the return-predicting ability of standard linear approaches to tree-based methods that allow us to replace these assumptions with data-driven counterparts. To address the question, we focus on commodities futures markets. These markets, with their information transparency, high liquidity, and low transaction costs, are close to the efficient market ideal that posits the unpredictability of returns.

Out of sample, we document three main findings: First, up to 3.74% of 1-month returns are predictable—a more than 10-fold increase from the standard linear approach. Second, moving from a simple linear risk factor approach to tree-based methods raises the in-sample (out-of-sample) predictability from 0.58% to 56.21% (from 0.33% to 3.74%). Third, tree-based portfolios systematically outperform their linear counterparts not only in terms of absolute returns but also in risk-adjusted terms.

Based on our findings, we contribute two important points to the debate on return predictability and market efficiency. First, much more return variation is predictable than previously suggested by linear methods. A possible interpretation is that markets may be more inefficient than previous work suggests. Second, the fact

that our results change quite dramatically when moving from a simple linear risk factor approach to tree-based methods hints at a different structure in the data. Therefore, an alternative way of testing new factors may be more appropriate: Rather than testing whether a factor adds explanatory power in isolation (or as an add-on to an oversimplified linear factor model), we should ask whether it adds explanatory power when it is added to an ensemble of existing factors taking into account multiway interactions and nonlinearities.

ACKNOWLEDGMENTS

We thank Frank Fabozzi, Ana-Maria Fuertes, David-Michael Lincke, Marcos López de Prado, and Friedrich Wetterling for useful comments. Annika Sophia Benecke provided excellent research assistance.

REFERENCES

Acemoglu, D. 2002. "Directed Technical Change." *The Review of Economic Studies* 69 (4): 781–809.

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen. 2013. "Value and Momentum Everywhere." *The Journal of Finance* 68 (3): 929–985.

Athey, S. *The Impact of Machine Learning on Economics.* Chicago: University of Chicago Press. 2019.

Athey, S., and G. W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725.

Bailey, D. H., and M. López De Prado. 2014. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality." *The Journal of Portfolio Management* 40 (5): 94–107.

Barro, R. J. 2006. "Rare Disasters and Asset Markets in the Twentieth Century*." *The Quarterly Journal of Economics* 121 (3): 823–866.

Bishop, C. M. *Neural Networks for Pattern Recognition.* Oxford: Oxford University Press, Inc. 1995.

Black, F. 1993. "Beta and Return." *The Journal of Portfolio Management* 20 (1): 8.

Breiman, L. 1996. "Bagging Predictors." *Machine Learning* 24 (2): 123–140.

——. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Brock, W., J. Lakonishok, and B. Lebaron. 1992. "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns." *The Journal of Finance* 47 (5): 1731–1764.

Criminisi, A., J. Shotton, and E. Konukoglu. 2012. "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning." *Foundations and Trends in Computer Graphics and Vision* 7 (2–3): 81–227.

De Bondt, W. F. M., and R. Thaler. 1990. "Do Security Analysts Overreact?" *American Economic Review* 80 (2): 52–57.

De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann. 1990. "Noise Trader Risk in Financial Markets." *Journal of Political Economy* 98 (4): 703–738.

DellaVigna, S., and J. M. Pollet. 2009. "Investor Inattention and Friday Earnings Announcements." *The Journal of Finance* 64 (2): 709–749.

Dietterich, T. G. 2000. "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization." *Machine Learning* 40 (2): 139–157.

Elton, E. J., M. J. Gruber, and C. R. Blake. 1996. "Survivorship Bias and Mutual Fund Performance." *The Review of Financial Studies* 9 (4): 1097–1120.

Fama, E. F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25 (2): 383–417.

Fama, E. F., and K. R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance* 47 (2): 427–465.

——. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

——. 2016. "Dissecting Anomalies with a Five-Factor Model." *The Review of Financial Studies* 29 (1): 69–103.

Farhi, E., and X. Gabaix. 2016. "Rare Disasters and Exchange Rates." *Quarterly Journal of Economics* 131 (1): 1–52.

Fayyad, U. M., and K. B. Irani. 1993. "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning." Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry, France, August 28–September 3, pp. 1022–1029.

Feng, G., S. Giglio, and D. Xiu. 2019. "Taming the Factor Zoo: A Test of New Factors." Working paper no. 25481, National Bureau of Economic Research Working Paper Series.

Fuertes, A. M., J. Miffre, and A. Fernandez-Perez. 2015. "Commodity Strategies Based on Momentum, Term Structure, and Idiosyncratic Volatility." *Journal of Futures Markets* 35 (3): 274–297.

Garcia, S., J. Luengo, J. A. Saez, V. Lopez, and F. Herrera. 2013. "A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning." *IEEE Transactions on Knowledge and Data Engineering* 25 (4): 734–750.

Grinblatt, M., and B. Han. 2005. "Prospect Theory, Mental Accounting, and Momentum." *Journal of Financial Economics* 78 (2): 311–339.

Gu, S., B, T. Kelly, and D. Xiu. 2020. "Empirical Asset Pricing via Machine Learning." *The Review of Financial Studies* 35 (5): 2223–2273.

Harvey, C. R., and Y. Liu. 2019. "False (and Missed) Discoveries in Financial Economics." *The Journal of Finance* (forthcoming).

Harvey, C. R., Y. Liu, and H. Zhu. 2016. "... and the Cross-Section of Expected Returns." *The Review of Financial Studies* 29 (1): 5–68.

Hirshleifer, D. 2015. "Behavioral Finance." *Annual Review of Financial Economics* 7 (1): 133–159.

Jorion, P., and W. N. Goetzmann. 1999. "Global Stock Markets in the Twentieth Century." *The Journal of Finance* 54 (3): 953–980.

Kothari, S. P., J. Shanken, and R. G. Sloan. 1995. "Another Look at the Cross-Section of Expected Stock Returns." *The Journal of Finance* 50 (1): 185–224.

Lo, A., and A. C. MacKinlay. 1990. "Data-Snooping Biases in Tests of Financial Asset Pricing Models." *The Review of Financial Studies* 3 (3): 431–467.

Longstaff, F. A. 1995. "How Much Can Marketability Affect Security Values?" *The Journal of Finance* 50 (5): 1767–1774.

López de Prado, M. L. *Advances in Financial Machine Learning.* Hoboken: Wiley, 2018.

MacKinlay, A. C. 1995. "Multifactor Models Do Not Explain Deviations from the CAPM." *Journal of Financial Economics* 38 (1): 3–28.

Mclean, R. D., and J. Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *The Journal of Finance* 71 (1): 5–32.

Moritz, B., and T. Zimmermann. 2016. "Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns." Working paper, 03, SSRN.

Moskowitz, T. J., Y. H. Ooi, and L. H. Pedersen. 2012. "Time Series Momentum." *Journal of Financial Economics* 104 (2): 228–250.

Novy-Marx, R. 2013. "The Other Side of Value: The Gross Profitability Premium." *Journal of Financial Economics* 108 (1): 1–28.

Rietz, T. A. 1988. "The Equity Risk Premium a Solution." *Journal of Monetary Economics* 22 (1): 117–131.

Schapire, R. E., and Y. Singer. 1999. "Improved Boosting Algorithms Using Confidence-Rated Predictions." *Machine Learning* 37 (3): 297–336.

Schwerhoff, G., and M. Stuermer. "Non-Renewable Resources, Extraction Technology, and Endogenous Growth." Working paper 2015(09), Federal Reserve Bank of Dallas, 2019.

Schwert, G. W. *Chapter 15 Anomalies and Market Efficiency,* Vol. 1, pp. 939–974. Elsevier, 2003.

Sharpe, W. F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *The Journal of Finance* 19 (3): 425–442.

Shleifer, A., and R. W. Vishny. 1997. "The Limits of Arbitrage." *The Journal of Finance* 52 (1): 35–55.

Struck, C., and E. Cheng. 2019. "Time-Series Momentum: A Monte Carlo Approach." *The Journal of Financial Data Science* 1 (4): 103–123.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Time-Series Momentum: A Monte Carlo Approach

Clemens Struck and Enoch Cheng

The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/4/103

ABSTRACT: *In this article, the authors develop a Monte Carlo backtesting procedure for risk premium strategies and employ it to study time-series momentum (TSM). Relying on time-series models, empirical residual distributions, and copulas, the authors address two key drawbacks of conventional backtesting procedures. They create 10,000 paths of different TSM strategies based on the S&P 500 and a cross-asset-class futures portfolio. The simulations reveal a probability distribution that shows that strategies that outperform buy and hold in sample using historical backtests may (1) exhibit tail risks and (2) underperform or outperform when out of sample. The results are robust to using different time-series models, time periods, asset classes, and risk measures.*

Enhancing Time-Series Momentum Strategies Using Deep Neural Networks

Bryan Lim, Stefan Zohren, and Stephen Roberts

The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/4/19

ABSTRACT: *Although time-series momentum is a well-studied phenomenon in finance, common strategies require the explicit definition of both a trend estimator and a position sizing rule. In this article, the authors introduce deep momentum networks—a hybrid approach that injects deep learning–based trading rules into the volatility scaling framework of time-series momentum. The model also simultaneously learns both trend estimation and position sizing in a data-driven manner, with networks directly trained by optimizing the Sharpe ratio of the signal. Backtesting on a portfolio of 88 continuous futures contracts, the authors demonstrate that the Sharpe-optimized long short-term memory improved traditional methods by more than two times in the absence of transaction costs and continued outperforming when considering transaction costs up to 2–3 bps. To account for more illiquid assets, the authors also propose a turnover regularization term that trains the network to factor in costs at run-time.*

Enhancing Risk-Adjusted Return Using Time Series Momentum in Sovereign Bonds

Gerhard Hambusch, KiHoon Jimmy Hong, and Ellenora Webster

The Journal of Fixed Income
https://jfi.pm-research.com/content/25/1/96

ABSTRACT: *This article studies an actively managed bond strategy based on time-series momentum in sovereign bond markets. The author assesses the performance of an active strategy and investigates diversification benefits in comparison with a passive buy-and-hold strategy when each strategy is combined with international equity indexes. The analysis provides evidence that the active strategy offers higher expected returns without increasing return volatility. Importantly, and in comparison with the passive strategy, the active strategy results in both significant return and diversification enhancements when combined with international equity indexes. Therefore, the author suggests that his active momentum strategy can serve fund managers as an alternative to common long-only passive bond strategies to enhance the risk-adjusted return of a combined portfolio of sovereign bonds and equities.*

Portfolio Selection Using Portfolio Committees

TSUNG-WU HO

TSUNG-WU HO
is a professor at the Graduate Institute of Global Business and Strategy of the College of Management at National Taiwan Normal University in Taiwan, R.O.C.
tsungwu@ntnu.edu.tw

KEY FINDINGS

- This article proposes a flexible and easy-to-implement committee approach to portfolio selection.
- This article defines an algorithm that proposes a score to select the best portfolio out of 250 augmented portfolios.
- In survival of the fittest, evidence from several datasets shows that the resulting combination portfolio overcomes the distributional uncertainty and exhibits superior annualized performance.

ABSTRACT: *The author proposes a committee approach to portfolio selection. Because each optimal portfolio is a combination of three basic elements—strategy, covariance matrix, and risk type—the author first augments the combination to 250 optimal portfolios at each estimation period. The author then defines a score to select the best portfolio to hold in the next period. Survival of the fittest, the superior performance of the combination portfolio, demonstrates that the committee approach to portfolio selection is not only effective but also easy to implement.*

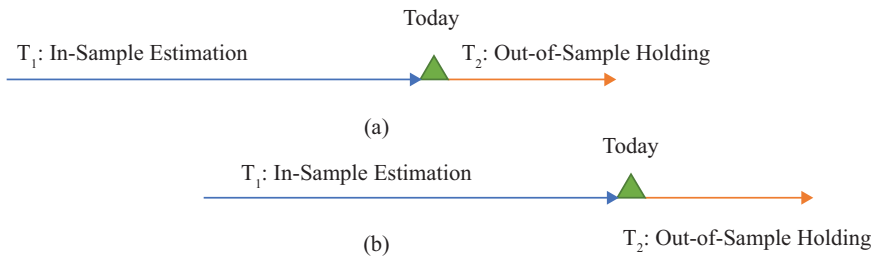
TOPICS: ***Portfolio theory, portfolio construction****

In machine learning, combining multiple models has been shown to boost predictive performance. Such combinations of models are often referred to as *committees* (Bishop 2009, Ch.14). The committee approach stands in contrast to standard portfolio selection, which provides a solution to an optimization problem within the confines of one model and moreover does not include a training or learning process. This article proposes a way to use committees in portfolio selection. Our basic methodology is simple: Because each optimal portfolio is a combination of three basic elements—strategy, covariance matrix, and risk type—we first augment the combination to 250 optimal portfolios at each estimation period, and then we define a score as an information criterion to select the best portfolio to hold in the next period. In survival of the fittest, our empirical results verify the effectiveness of this method.

The most popular portfolio optimization frameworks in finance are what we call *risk diversification*, *risk parity*, and *risk optimal*. The risk-diversification portfolio is a mean–variance-type portfolio, and the risk-optimal portfolio is an expected shortfall or conditional value-at-risk (CVaR)-type portfolio. For risk-diversification portfolios, greater

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

EXHIBIT 1
The Out-of-Sample Problem of Portfolio Optimization



benefits can be obtained from diversification when the assets included in the portfolio are less correlated. In the seminal work of Markowitz (1952), an efficiently diversified portfolio is obtained through mean–variance optimization. However, researchers have indicated that, when using the real data, portfolio selection is not robust in two respects:

1. First, the mean–variance model is sensitive to assumed normality. Tobin (1958) and Chamberlain (1983) indicated that the mean–variance framework appropriately captures the trade-off between risk and return only if the distribution of returns is elliptically symmetric. Fama (1965) and Kon (1984) indicated that, once the real data are not elliptically symmetric, mean and variance will not be sufficient statistics for investors to make optimal portfolio decisions.
2. The second respect is the presence of distributional uncertainty or ambiguity: Results from the in-sample estimation period are sensitive to the ambiguity of the out-of-sample holding period. The holding period ambiguity is associated with distributional uncertainty. Uppal and Tan (2003) and Simonian and Davis (2011) formally defined uncertainty as ambiguity in the sense of Ellsberg (1961) and analyzed its effect on portfolio selection.[1] Jagannathan and Ma (2003) explained why there is a possible trade-off between specification error and sampling error.

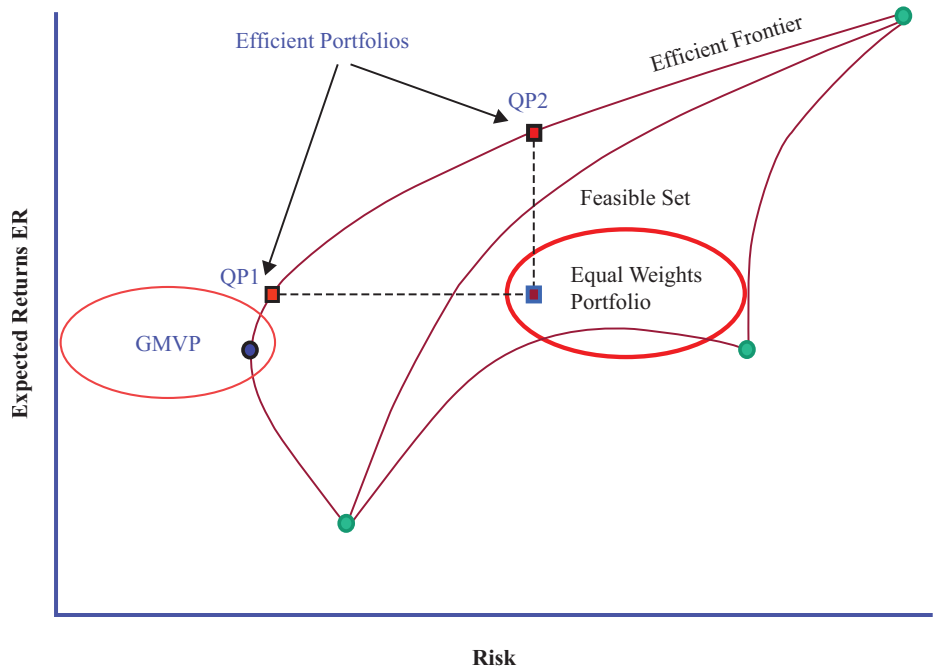
Exhibit 1 illustrates the problem with risk-diversification portfolios. Assuming an investment horizon $T_1 + T_2$, standard portfolio selection solves the optimization problem by combining a given covariance matrix, risk type, and portfolio strategy at T_1 and then uses the solved weight vector to build a portfolio over T_2 to obtain the expected returns and risk. Because T_2 is totally unknown to decision makers at T_1, the distributional uncertainty of T_2 challenges the robustness of this selection methodology. That is, if the solved optimality of T_1 substantially deviates from the optimality of T_2, results may yield significant prediction errors and financial loss.

As a result of the foregoing model risk, simply selecting another optimization framework may not be sufficient to address the problem of out-of-sample uncertainty. According to machine learning, training is a good method to boost robustness. We then design a mechanism to train portfolio selection to improve robustness. The idea is simple: Because each optimal portfolio is a combination of three basic elements (strategy, covariance matrix, and risk type), we first augment the combination to 250 optimal portfolios at T_1. Second, we propose a score to select the winner out of 250 combinations and hold over at T_2. The combination of the three elements switches over time. Furthermore, we use two datasets to assess the effectiveness of the combination portfolio: the component stocks on the S&P 100 and 31 exchange-traded funds (ETFs). Our backtesting results reveal that the combination portfolio substantially outperforms other portfolios.

This article is organized as follows: The second section introduces the mean–variance (MV) diversification and risk-optimal portfolios. The third section explains the content of the combination portfolio. The fourth section discusses the measures of robust covariance. The fifth section presents the empirical results. The sixth section

[1] Other examples of ambiguity-sensitive optimization frameworks include those by Rustem and Howe (2002); Garlappi, Uppal, and Wang (2007); Ben-Tal, El Ghaoui, and Nemirovski (2009); Fabozzi et al. (2007); Simonian and Davis (2010); and Simonian (2011, 2019).

EXHIBIT 2
The Mean–Variance Portfolio Frontier



Note: GMVP is the global minimum variance portfolio.

compares implied expected returns of the combination portfolio with the realized expected returns. The seventh section concludes the article.

OPTIMIZATION STRATEGIES IN PORTFOLIO LITERATURE

Exhibit 2 depicts the efficient frontier from classical finance theory. On the efficient frontier, we see the risk and return profiles for various portfolios. Efficient portfolios are those that provide the highest return for their level of risk.

Given this background, we introduce the portfolio strategies.

Diversification Portfolios: The Mean–Variance Type

Here we introduce the four basic portfolio strategies.

1. *Risk minimization with an expected return target*. In this portfolio, we choose an expected portfolio return target that functions as a constraint on an optimization that seeks to minimize portfolio risk. This portfolio is represented by QP1 in Exhibit 2. We can implement the optimization of this portfolio using quadratic programming:

$$\min_{w} \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\Omega\mathbf{w}$$
$$s.t.$$
$$\mathbf{Erw} = \bar{\mathbf{R}}$$
$$\mathbf{w}^{\mathrm{T}}\mathbf{1} = 1 \qquad (1)$$

where $\mathbf{\Omega}$ denotes the covariance matrix, $\mathbf{w}$ represents the weight vector, and $\mathbf{r}$ denotes the time series of asset returns. $\bar{\mathbf{R}}$ denotes the level of target portfolio returns assumed to be the time-series average of the weighted portfolio.

2. *Returns maximization, given a risk target*. In the second strategy, we assume a specific risk target, which serves as a constraint on an optimization that seeks to maximize return. This portfolio is

represented by QP2 in Exhibit 2. As in the case of Equation 1, we can use quadratic programming to solve the optimization.

$$\begin{aligned} \max_{w} \quad & \mathbf{E}R = \mathbf{E}(\mathbf{rw}) \\ s.t. \quad & \mathbf{w}^{\mathbf{T}}\Omega\mathbf{w} = \bar{\sigma} \\ & \mathbf{w}^{\mathbf{T}}\mathbf{1} = 1 \end{aligned} \tag{2}$$

where the expected return **ER** is the weighted returns.

3. *Sharpe ratio maximization*. The third strategy searches for the maximal Sharpe ratio along the frontier, namely, the ratio of *y*-axis returns and *x*-axis risk. This strategy is also known as the *tangency portfolio*. The Sharpe ratio maximization strategy solves the following objective function:

$$\begin{aligned} \max_{w} \quad & \frac{\mathbf{Erw} - r_F}{\mathbf{w}^{\mathbf{T}}\Omega\mathbf{w}} \\ s.t. \quad & \mathbf{w}^{\mathbf{T}}\mathbf{1} = 1 \end{aligned} \tag{3}$$

4. *Global minimum variance*. As indicated by Exhibit 2, the point with the smallest risk on the efficient frontier is the GMVP, which represents just the minimum risk point on the efficient frontier, which is the vertical tangent line to the left-hand curve.

The set of weight is solved analytically as follows:

$$w = \frac{\Omega^{-1}\mathbf{1}}{\mathbf{1}^{\mathbf{T}}\Omega^{-1}\mathbf{1}}$$

Risk-Optimal Portfolio: CVaR and MAD Types

The risk-optimal portfolio constructs the portfolio whose mean return is associated with tail risk. Hence, the investors ask questions such as how the portfolio weights should be chosen such that the resulting 99% CVaR is 3.5%. The optimization problem of mean-CVaR portfolio selection is formally expressed as follows:

$$\begin{aligned} \min_{w} \quad & CVaR_{5\%}(w) \\ s.t. \quad & \mathbf{Erw} = \bar{\mathbf{R}} \\ & \mathbf{w}^{\mathbf{T}}\mathbf{1} = \mathbf{1} \end{aligned} \tag{4}$$

In addition, the mean absolute deviance (MAD) frontier uses absolute deviation as a risk measure. Typically, the deviation is derived from the central value, being construed as some type of an average, most often the median or the mean of the dataset.

The Combination Portfolio

Based on the foregoing four strategies, we explain the methodology behind the construction of 250 candidate portfolios and assigning scores to determine the winner.

We begin with 25 portfolio strategies. Optimization of a diversified portfolio involves the specification of four solution strategies as objective functions, as mentioned previously: minrisk, maxreturn, tangency, and GMVP. Together with the basic four strategies, the other 21 solutions are versions of the Sharpe ratio–based tangency portfolio, referred to as the *augmented tangency portfolios*.

The conventional solution to the tangency portfolio is to solve the maximal standard Sharpe ratio along the efficient frontier. Consider 49 industry portfolios as shown in Exhibit 3. Once the risk and return of each industry portfolio are plotted, we obtain an efficient frontier. Any point on the frontier is a portfolio with a pair of risk and expected return values; accordingly, the corresponding value of the Sharpe ratio can be naturally computed (indicated by the values marked by the right-hand-side scale bar, ranging from 0.0428 to 0.0913).

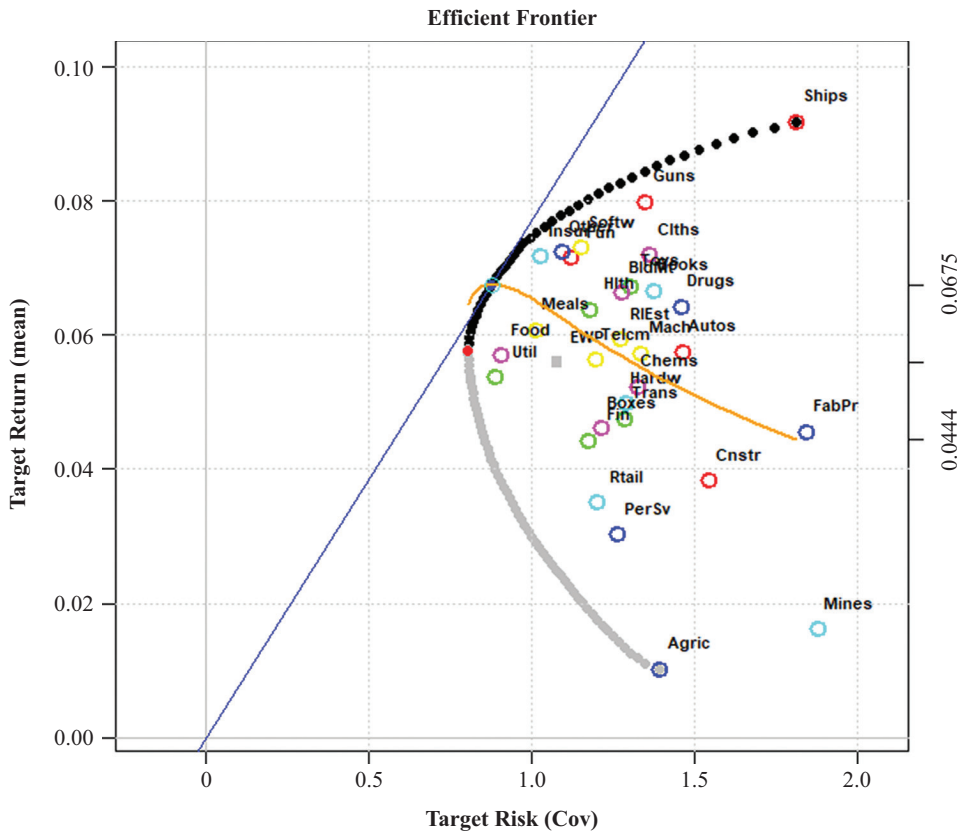
Under the MV diversification framework, we augment the Sharpe-style portfolio by applying more performance metrics. In our experiment, we use 21 performance ratios to construct 21 tangency portfolios. This leaves us with 25 solution strategies for portfolios. The 21 ratios are listed in Exhibit 4.

In Exhibit 4, some metrics' denominators are adjusted by VaR (ES) (e.g, SharpeRatioVaR and SharpeRatioES). We do not describe the formal details of most of these metrics in Exhibit 4 because they are commonly known. However, we briefly explain the AdjustedSharpeRatio, which is designed to capture the statistical properties of skewness and kurtosis. It is formally described as follows:

$$SR \cdot \left[1 + SR \cdot \left(\frac{skewness}{6} \right) - SR^2 \cdot \left(\frac{kurtosis - 3}{24} \right) \right]$$

where *SR* denotes the annualized Sharpe ratio.

EXHIBIT 3
The Solution of the Tangency Portfolio



Every mean–variance-type optimization requires a covariance matrix. We select eight covariance matrixes as inputs (described in detail in the fourth section). Taken together, the number of solution strategies and covariance matrixes gives us 200 possible portfolios, which we summarize in Exhibit 5.

Moreover, given 25 solution strategies, we also consider two risk-optimal portfolios: CVaR and MAD. When combined with the 25 solution strategies, we have an additional 50 solutions, giving us a total of 250 possible portfolios.

Computation of the Score

We evaluate each portfolio by assigning a score. We compute two basic performance ratios: the annualized Sharpe ratio and the adjusted Sharpe ratio. In Equation 5, $ER_t^{(a)}$ denotes the annualized expected returns, and r_F denotes the risk-free rate. The risk measures in the denominator are annualized measures. The risk in the denominator in each of the latter metrics is measured by annualized standard deviation, VaR, and CVaR(ES).

$$SR_1 = \frac{ER_t^{(a)} - r_F}{\hat{s}},\ SR_2 = \frac{ER_t^{(a)} - r_F}{VaR},\ SR_3 = \frac{ER_t^{(a)} - r_F}{ES} \tag{5}$$

The skewness- and kurtosis-adjusted annualized Sharpe ratio is defined as

$$adjSR_i = SR_i \cdot \left[1 + SR_i \cdot \left(\frac{skewness}{6}\right) - SR_i^2 \cdot \left(\frac{kurtosis - 3}{24}\right)\right] \tag{6}$$

where $i = 1, 2, 3$. Overall, the score is defined as the average as follows:

$$score = \frac{\sum_{i=1}^{3} SR_i + \sum_{i=1}^{3} adjSR_i}{6} \tag{7}$$

EXHIBIT 4
The 21 Augmented Tangency Portfolios under MV Diversification

Symbols	Implications
BernardoLedoitRatio	Bernardo–Ledoit ratio
SkewnessKurtosisRatio	Skewness–kurtosis ratio
CalmarRatio	Calmar ratio
SterlingRatio	Sterling ratio
BurkeRatio	Burke ratio
KellyRatio	Kelley ratio
MartinRatio	Martin ratio
PainRatio	Pain ratio
SharpeRatioVaR	Sharpe ratio adjusted by VaR
SharpeRatioES	Sharpe ratio adjusted by CVaR
SharpeRatio.annualized	Annualized Sharpe ratio
SharpeRatio.annualizedVaR	Annualized Sharpe ratio adjusted by VaR
SharpeRatio.annualizedES	Annualized Sharpe ratio adjusted by CVaR
AdjustedSharpeRatio	SharpeRatio.annualized adjusted by skewness and kurtosis
AdjustedSharpeRatioVaR	Adjusted SharpeRatioVaR
AdjustedSharpeRatioES	Adjusted SharpeRatioES
AppraisalRatio	Appraisal ratio
InformationRatio	Information ratio
TreynorRatio	Treynor ratio
OmegaSharpeRatio	Omega-Sharpe ratio
SortinoRatio	Sortino ratio

EXHIBIT 5
Universe of Possible Portfolios

8 Covariance Matrixes	25 Frontier Solution Strategies				
	Minrisk	Maxreturn	GMVP	Tangency	21 Augmented Tangency
Sample	1	1	1	1	21
GoldSach	1	1	1	1	21
LW	1	1	1	1	21
ShrinkCC	1	1	1	1	21
Student *t*	1	1	1	1	21
Shrinkage	1	1	1	1	21
Slpm	1	1	1	1	21
Spearman	1	1	1	1	21

The choice of both ratios is completely data driven; the technical details are explained in the Appendix. We proceed to describe the mechanics of our algorithm.

The Algorithm

Our first step is to select the horizons to estimate data. We consider 1 year (250 trading days) to estimate the performance of our 250 portfolio combinations, select the best one to hold for 1 month (22 trading days), and compute the out-of-sample performance. We repeat the exercise by moving forward 1 month at a time to rebalance the portfolio.

Next, for each iteration, we back test a single portfolio over time, where we compare its performance with the 250 combination portfolios and a benchmark, the

EXHIBIT 6
The Algorithm for Combination Portfolio

Step 1. Using 12-month (18-month) daily data, compute 250 optimal portfolios.
Step 2. For each portfolio, compute a mean score of six performance metrics.
Step 3. Select the best portfolio and hold it for 1 (3) month(s).
Step 4. Rebalance and compute the transaction fee.
Step 5. Iterate.

GSPC index. For simplicity, we set both the daily risk-free rate (r_F) and minimal acceptance returns (MAR) to be zero. Moreover, the transaction cost is set to be 0.25% and is counted twice for both buy and sell. For simplicity, the study does not consider transaction costs ex ante.

Exhibit 6 summarizes the algorithm for computing a combination portfolio.

Eight Covariance Estimators

In addition to a portfolio strategy, diversification must estimate a proper covariance matrix. The elements of a sample covariance matrix are usually estimated imprecisely. Several solutions have been suggested. The first approach is to impose more structure on the covariance matrix to reduce the number of parameters to be estimated—for example, constant correlation (Ledoit and Wolf 2004). The second approach is to use shrinkage estimators, shrinking the sample covariance matrix toward the single index, as proposed by Ledoit and Wolf (2003) and Candelon, Hurlin, and Tokpavi (2012). The third approach is to estimate the covariance matrixes by robust statistics. These include the multivariate Student t distribution, Spearman's rank estimator, and Bayesian shrinkage.

This study initializes eight covariance matrixes. In addition to the conventional sample covariance, the others are briefly explained in Appendix B.

Empirical Results

We have back tested our methodology on a selection of datasets to check the performance of the combination portfolio. These include the DJ index, French's 49 Industry Portfolio, Taiwan 50 Index, and China's CSI. Here we report results relating to two sets of securities: the S&P 100 component stocks and a basket of ETFs.[2] Each dataset is downloaded from Yahoo Finance, and the daily returns are computed by differencing the log of the adjusted close price.

S&P 100. For the S&P 100, approximately 83 stocks have no missing observations during the estimation sample period, ranging from May 15, 1997, to May 21, 2019. Exhibits 7–9 present the results. For brevity, we plot the top 30. The initial conclusion is that the combination portfolio performs best when the holding period is 1 month. Longer estimation horizons (e.g., 18 months) improve performance. By contrast, a longer holding period reduces performance, which may be a result of out-of-sample uncertainty.

We further examine the impacts of the holding period in terms of the annualized Sharpe ratio. We pick the two best portfolios from the pool without the combination and the GSPC index and then put them together with the combination and benchmark portfolios. We thus compare the magnitude of transaction fee and maximal drawdown with respect to different holding periods.

The previous annualized measures are based on the 3 years from 2015 to 2018.[3] To this end, we compute the rolling estimates of the one-step-ahead annualized net returns and Sharpe ratios over 3-year windows. With this design, we can measure the performance when an investor selects this portfolio at any point in time and holds it for 3 years. We display the results in Exhibit 9. As the exhibit shows, the annualized Sharpe ratio provides evidence for the strength of the described strategy.

ETFs. ETFs are, in general, index-tracking vehicles; consequently, the conventional four portfolio selection strategies do not further improve their portfolio performance, as Exhibit 10 indicates. Hence, ETFs are good candidates to test our algorithm for its ability to construct a combination portfolio that can outperform index funds. This dataset comprises the daily returns of 31 ETFs from July 31, 2002 to May 21, 2019, derived from Yahoo Finance. The symbols for the ETFs are XLB, XLE, XLF, XLP, XLI, XLU, XLV, XLK, XLY, RWR, SPY, EWJ, EWG, EWU, EWC, EWY, EWA, EWH, EWS, IYZ, EZU, IYR, EWT, EWZ, EFA, IGE, EPP, LQD, SHY, IEF, and TLT. In this dataset, SPY is

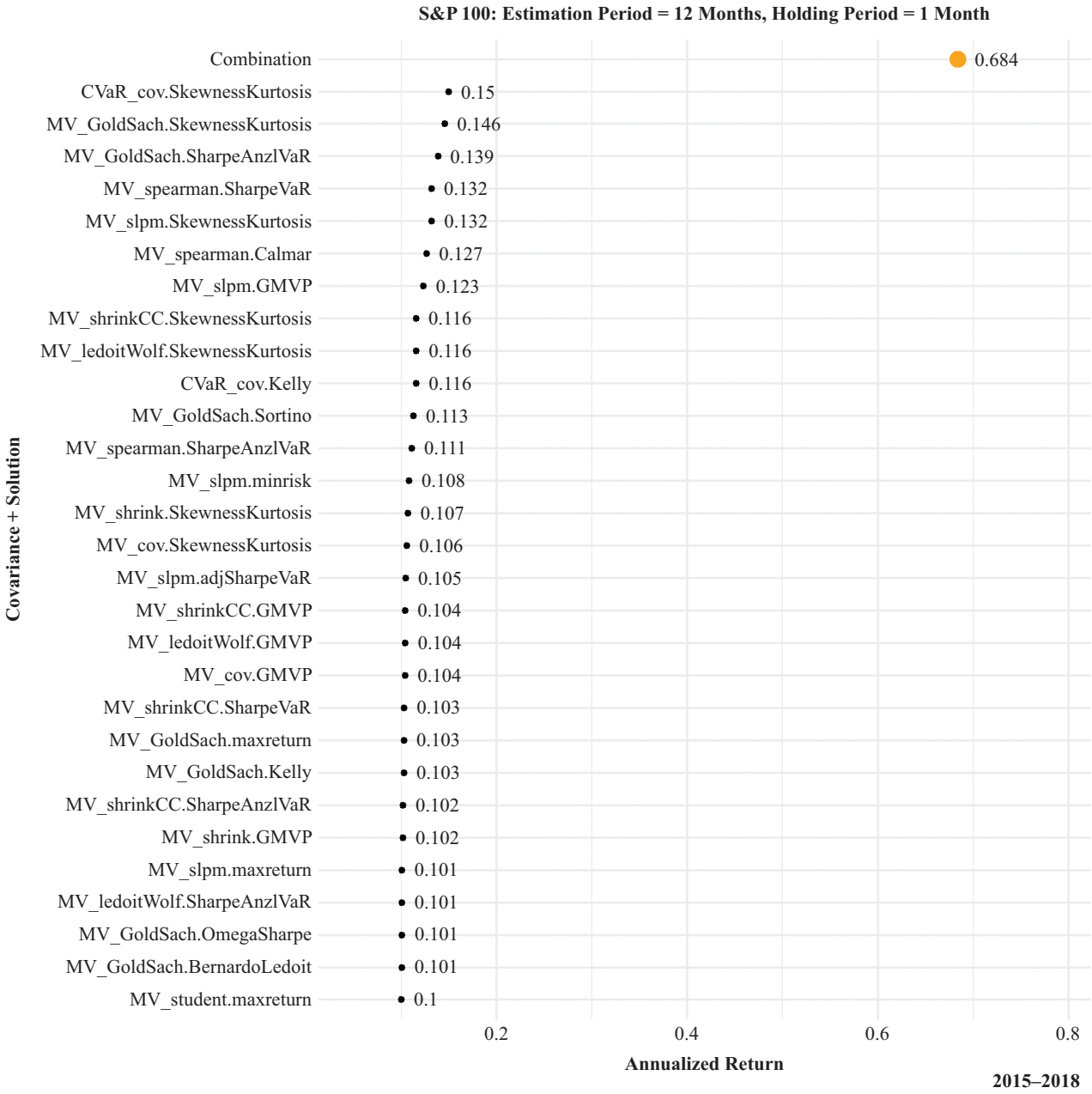
[2] More results are available upon request.

[3] We note that the good performance of the combination portfolio is robust and not sample dependent.

Exhibit 7

Portfolios Ranked by Annualized Sharpe Ratio: S&P 100

Panel A: Annualized Return, 12-Month Estimation Period, 1-Month Holding Period

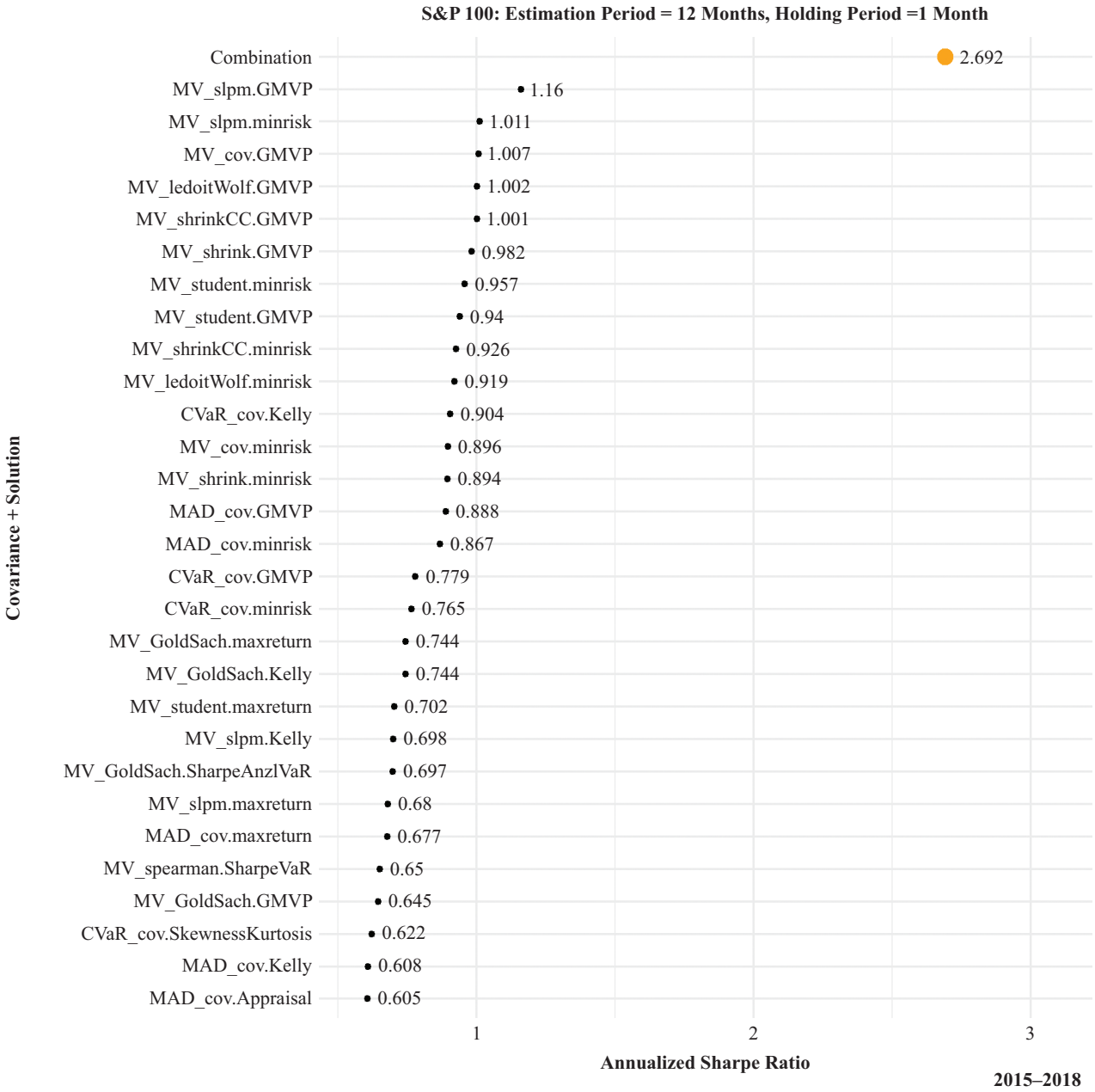


(continued)

EXHIBIT 7 *(continued)*

Portfolios Ranked by Annualized Sharpe Ratio: S&P 100

Panel B: Annualized Sharpe Ratio, 12-Month Estimation Period, 1-Month Holding Period

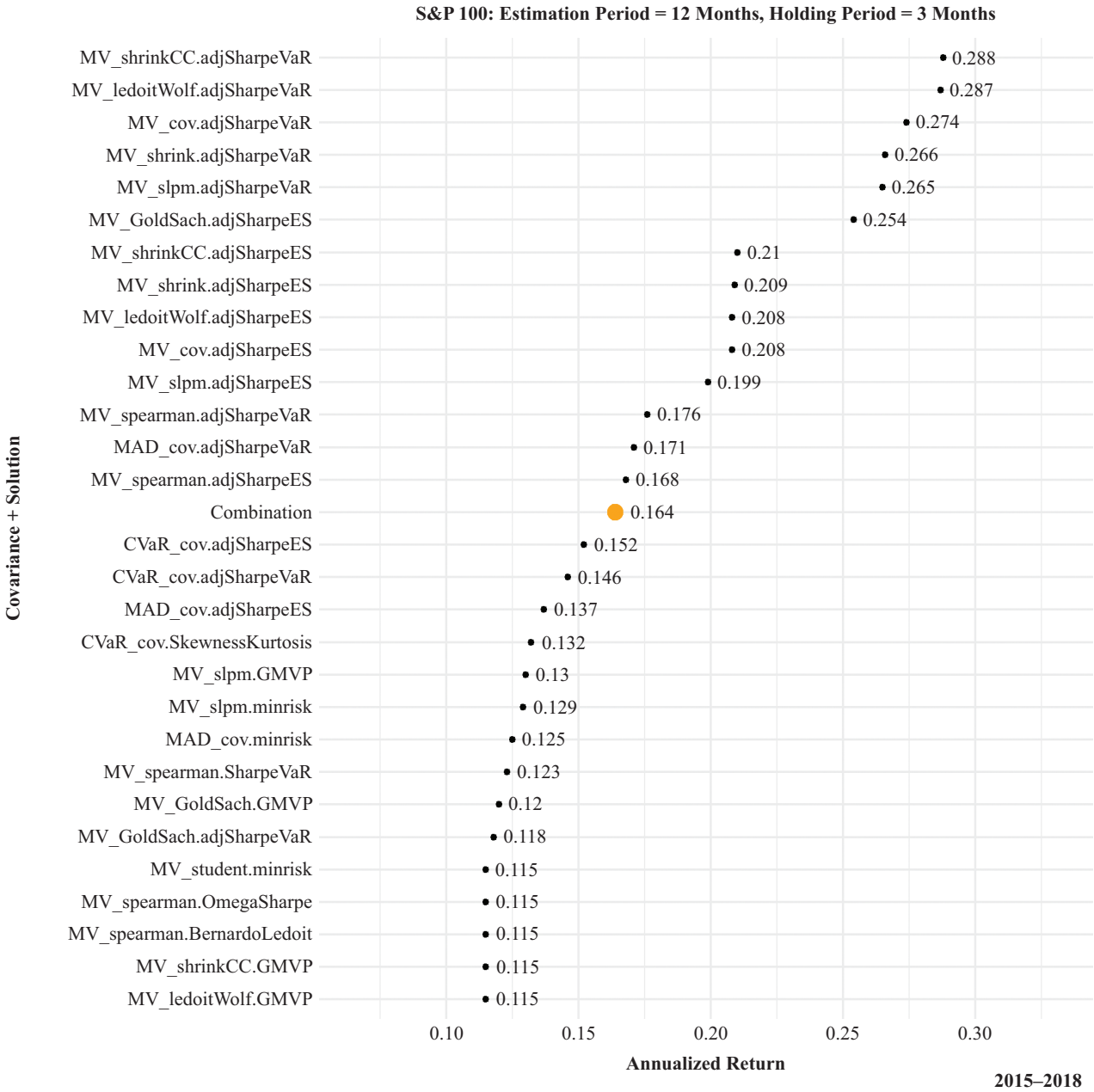


(continued)

EXHIBIT 7 *(continued)*

Portfolios Ranked by Annualized Sharpe Ratio: S&P 100

Panel C: Annualized Return, 12-Month Estimation Period, 3-Month Holding Period

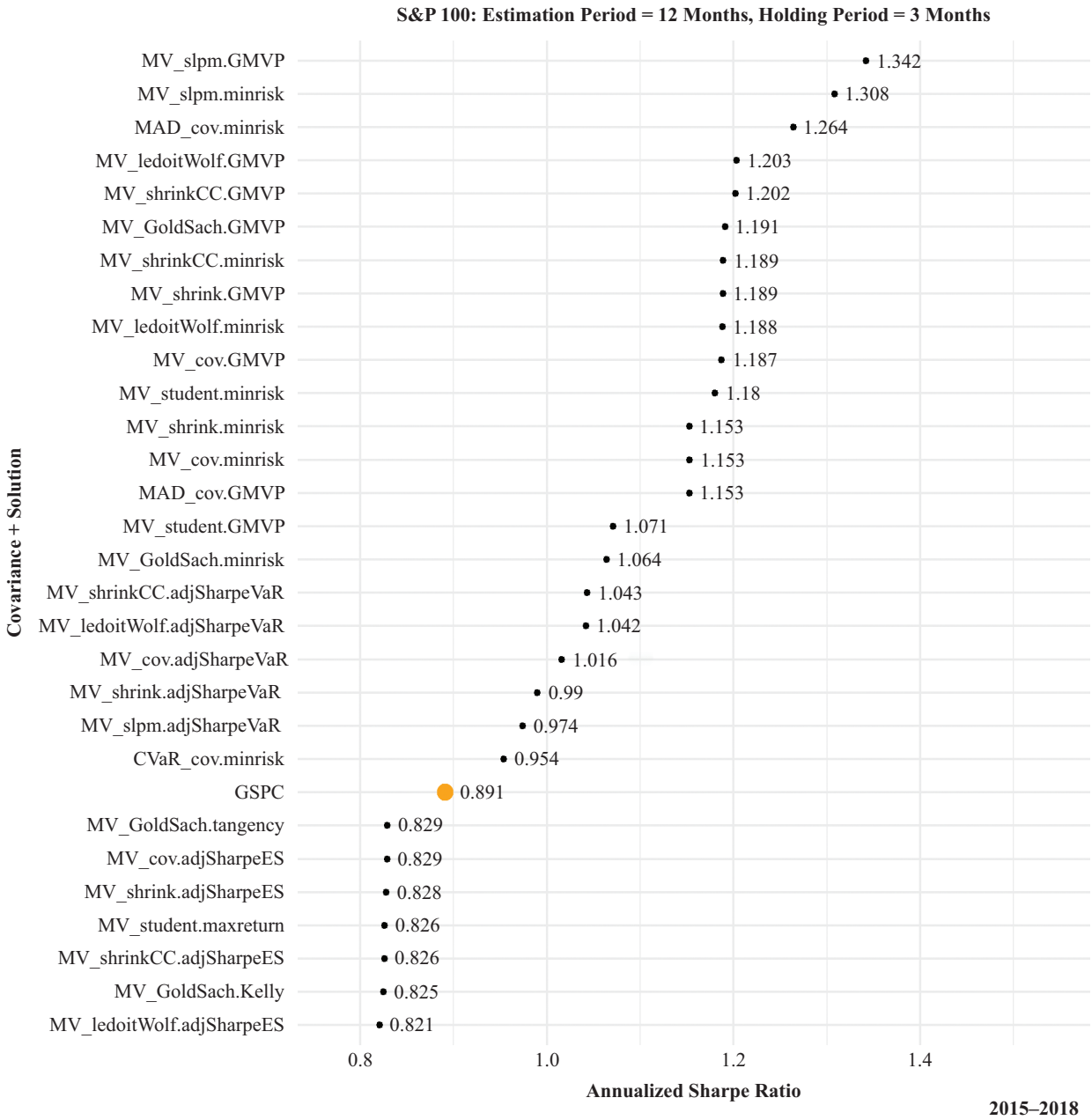


(continued)

EXHIBIT 7 *(continued)*

Portfolios Ranked by Annualized Sharpe Ratio: S&P 100

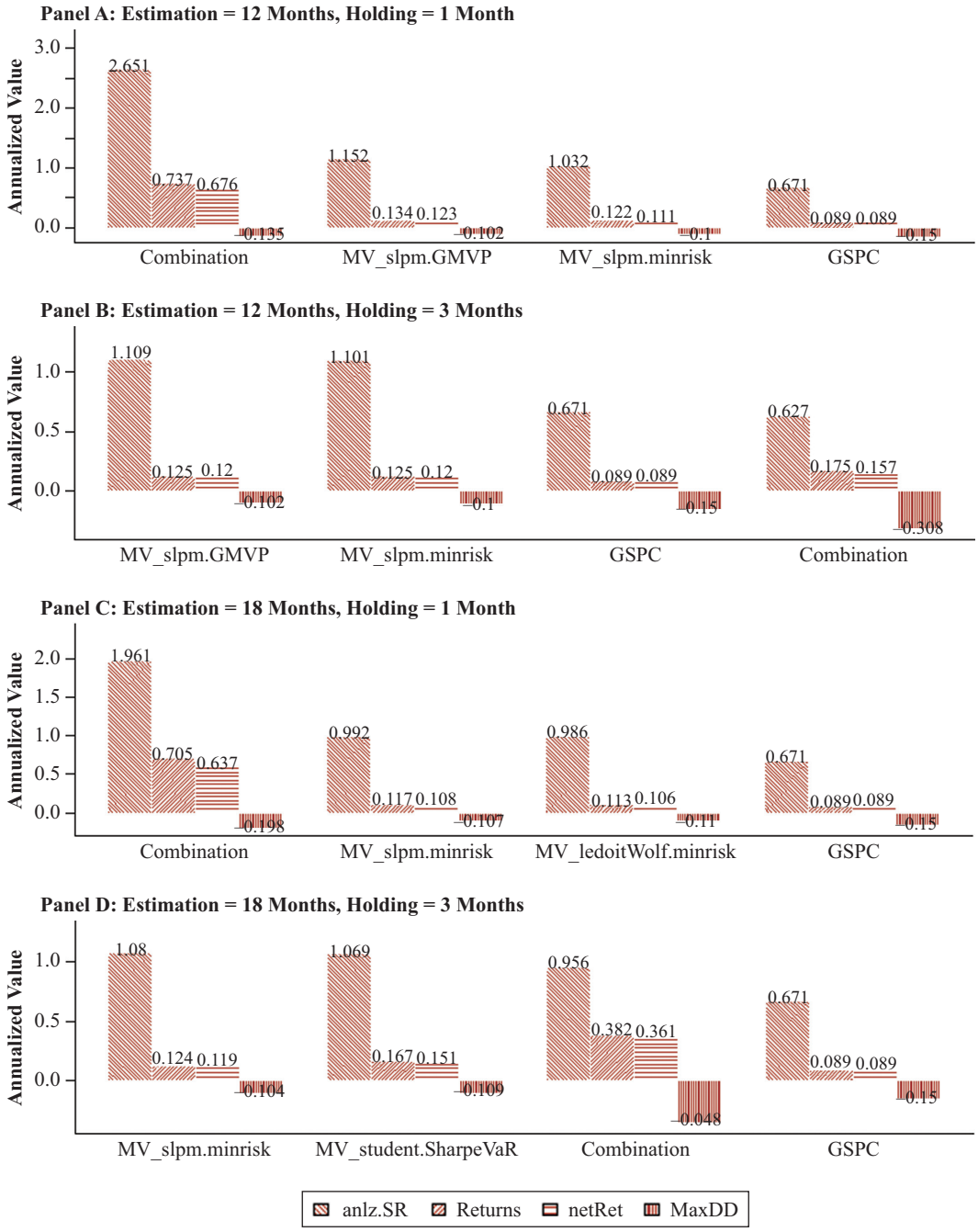
Panel D: Annualized Sharpe Ratio, 12-Month Estimation Period, 3-Month Holding Period



Notes: To illustrate performance, we compute the final 3-year (2015–2018) annualized Sharpe ratio of returns net of the transaction fee. Panels A and B graphically summarize the annualized returns and Sharpe ratio of 12 months and 1 month, which indicate that the combination portfolio outperforms other models, including the GSPC index returns. However, Panels C and D indicate that, as the holding period increases up to 3 months, the performance deteriorates: The combination portfolio fails to enter the top 30 with regard to the annualized Sharpe; hence, the top 30 does not cover it.

EXHIBIT 8

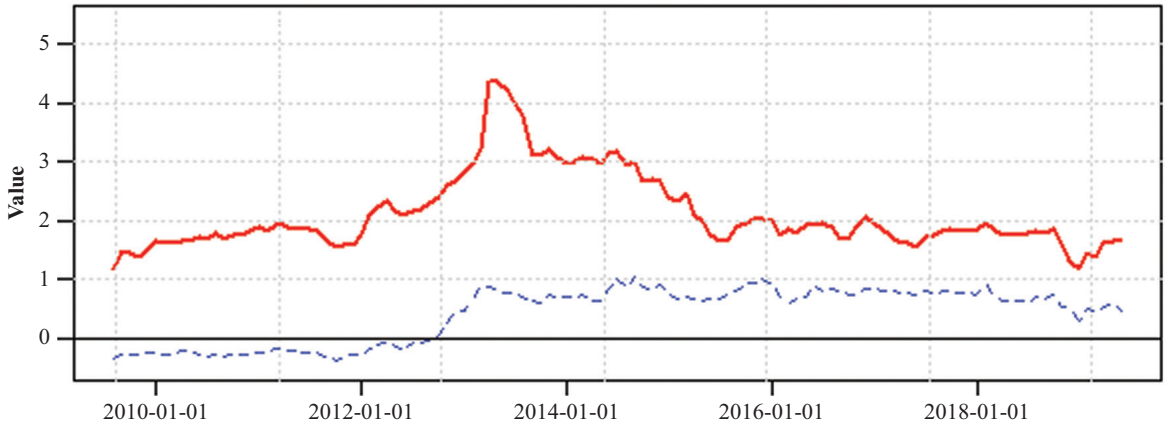
Comparison of Transaction Fee and Maximal Drawdown: S&P 100



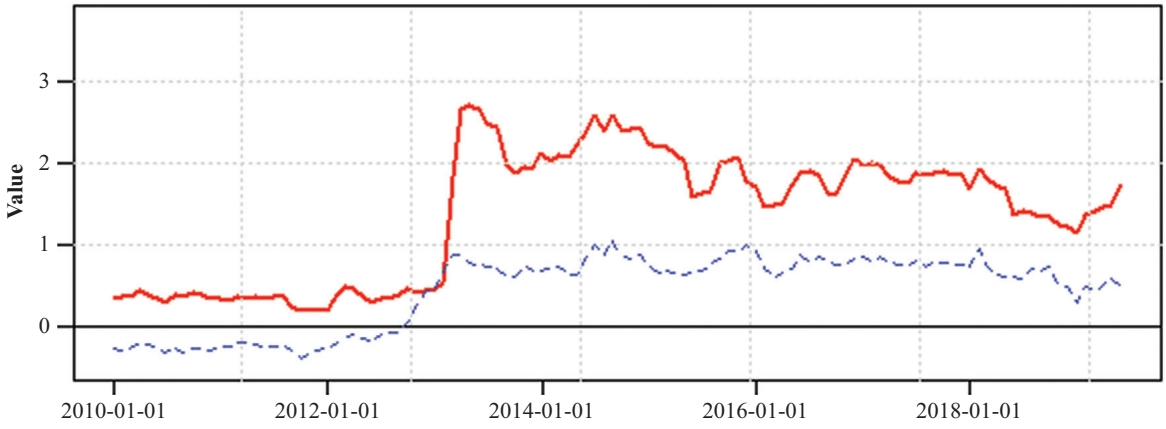
Notes: anlz.SR is the annualized Sharpe ratio, and MaxDD is the maximal drawdown. Panels A and B compare 12 months/1 month and 12 months/3 months: First, the net return of the combination portfolio is 68.1%, which is eight times larger than the annualized GSPC returns of 8.1%. However, the performance reversed when the holding period increased by 3 months. Second, in Panel A, the transaction fee of the combination portfolio is also large. The transaction fee is roughly 6% (74.3% − 68.1%), approximating the annualized GSPC index returns. Finally, the maximal drawdown of the combination portfolio is −13.5%, and the GSPC index is −15%, which seems comparable. Moreover, when the holding period is increased to 3 months, the maximal drawdown drops to −30.8%, and the overall performance worsens as well. Panels C and D compare 18 months/1 month and 18 months/3 months. When the estimation period is 18 months, the 1-month holding period indicates that the net return of the combination portfolio is 62%, which is 7.5 times larger than the annualized GSPC index returns. The performance reverses when the holding period increases to 3 months. The transaction fee of 18 months/1 month is approximately 7% (68.7% − 62%), and the maximal drawdown also becomes larger by −39%.

EXHIBIT 9
Rolling Estimated Compared: S&P 100

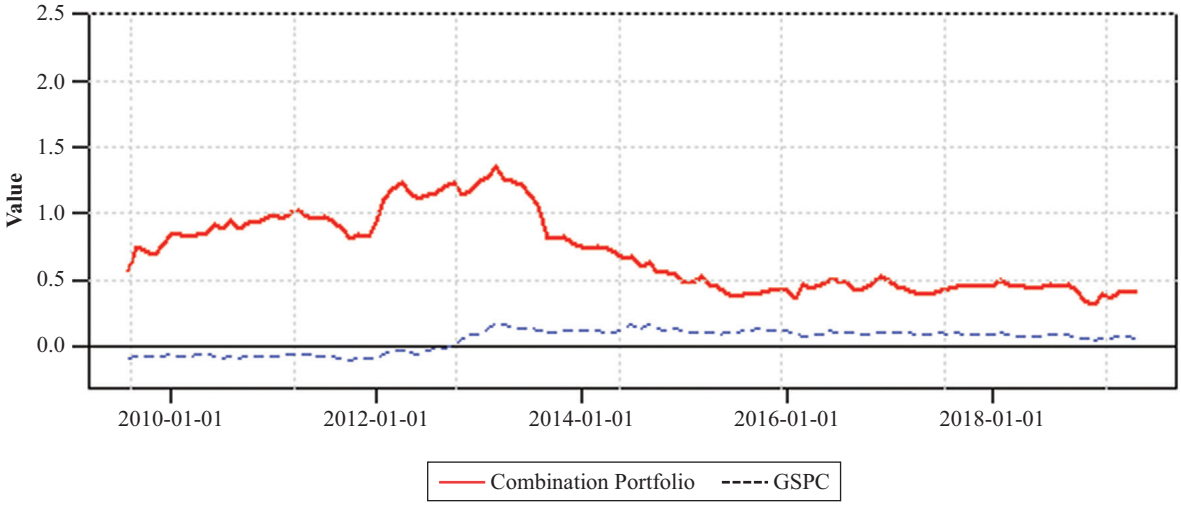
Panel A: Annualized Sharpe Ratio: Estimation = 12 Months, Holding = 1 Month



Panel B: Annualized Sharpe Ratio: Horizon = 18 Months, Holding = 1 Month



Panel C: Annualized Returns: Estimation = 12 Months, Holding = 1 Month



(continued)

EXHIBIT 9 *(continued)*

Rolling Estimated Compared: S&P 100

Panel D: Annualized Returns: Horizon = 18 Months, Holding = 1 Month

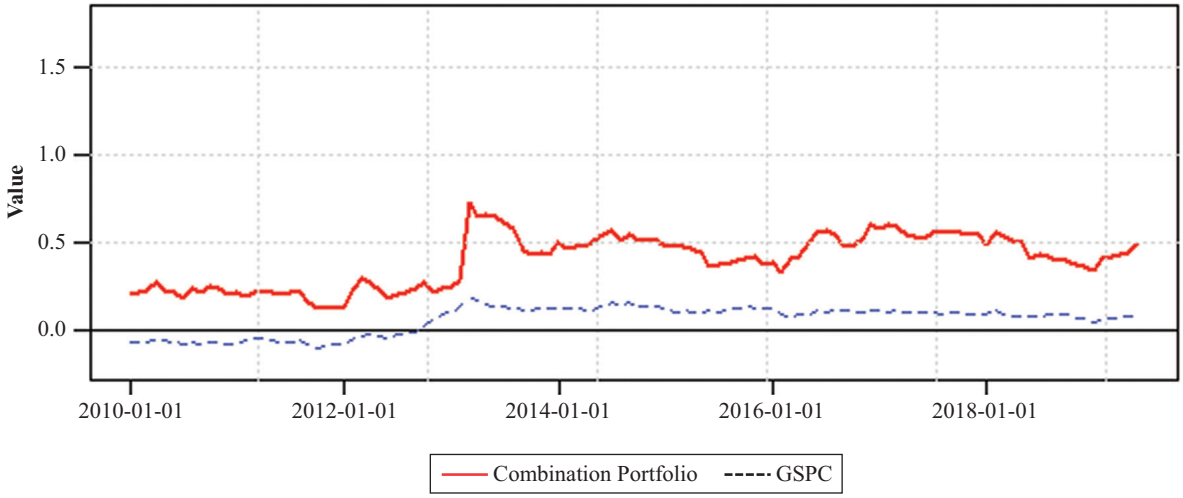
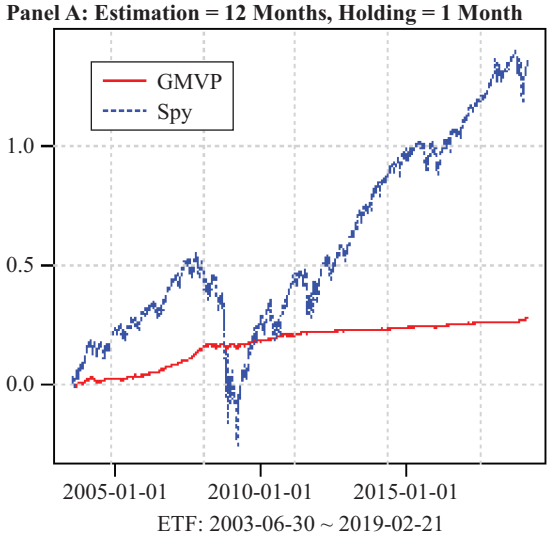


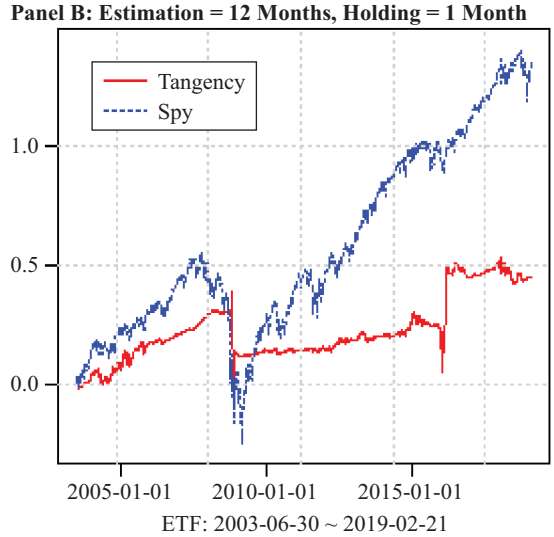
EXHIBIT 10

Backtesting Result of Four Basic Strategies for ETF

Panel A: Estimation = 12 Months, Holding = 1 Month



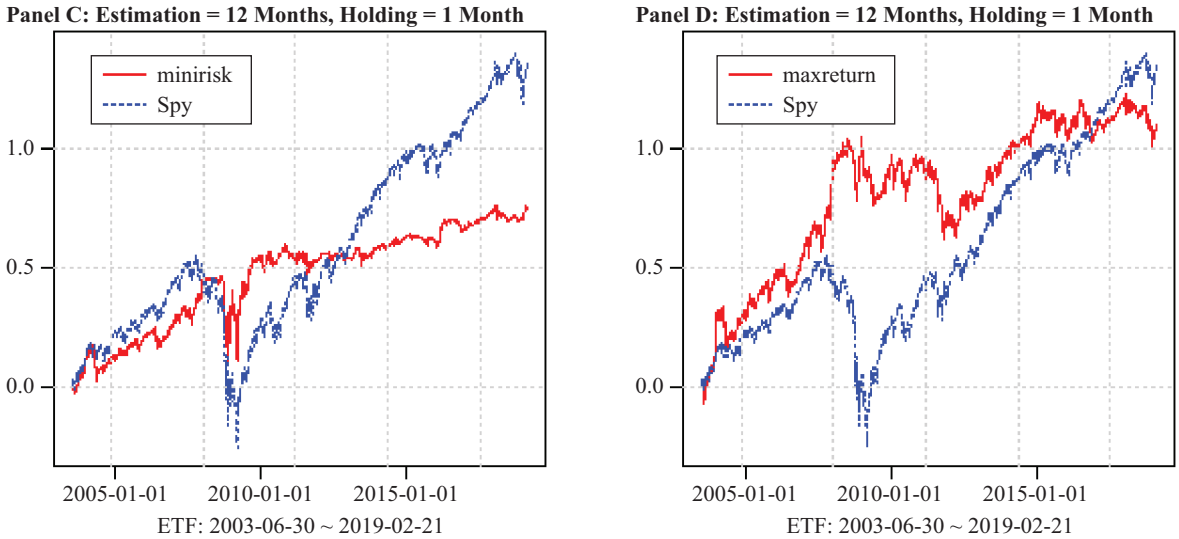
Panel B: Estimation = 12 Months, Holding = 1 Month



(continued)

EXHIBIT 10 *(continued)*

Backtesting Result of Four Basic Strategies for ETF



used as the benchmark because it is the best-recognized and oldest ETF. Exhibits 11–13 present our results. They show that the combination portfolio outperforms not only SPY but also other ETFs. Rolling estimates of Exhibits 12 and 13 further verify the results. The empirical evidence of the combination portfolio indicates that, in financial data science, training by portfolio theory is important.[4]

Simulated Data

In this section, we evaluate, via Monte Carlo simulation, the performance of the portfolio committee method. We use standard asset pricing theory and assume the risk-free rate is zero:

$$R_{it} = \beta R_{mt} + e_{it} \quad \text{for } i = 1 \ldots N \text{ and } t = 1 \ldots T$$

We first construct market returns R_{mt} from a cross section of 600 × 500 multivariate assets returns. The 600 × 500 random samples are generated from a multivariate generalized hyperbolic distribution, and the data-generating process imposes a skewed distribution of increasing negative skewness (Azzalini 1985). The random variables are scaled and shifted such that the means of the asset returns are

$$\mu_j = 0.015 + 0.025j \qquad j = 1, 2 \ldots 500$$

To reflect the positive risk–return trade-off property for market returns, we modify the technique of Weiqiang et al. (2012) and choose the skewness parameters to be $\gamma_j = -0.024975j$ such that assets with higher mean returns have higher negative skewness.[5]

Next, specific asset returns R_{it} are generated by sampling systematic risk β from the interval (0, 1), with random Gaussian error terms e_{it}. We consider the portfolio selection problem by simulating $N = \{10, 30\}$ asset returns with $T = 600$ observations, roughly a 2-year horizon. Finally, we compute the out-of-sample returns associated with the portfolio committee. The procedure is repeated 10,000 times for each N, and the overall performance is evaluated by averaging 10,000 annualized Sharpe ratios. Exhibit 14 lists the first 10 portfolios. As the exhibit shows, the combination portfolio

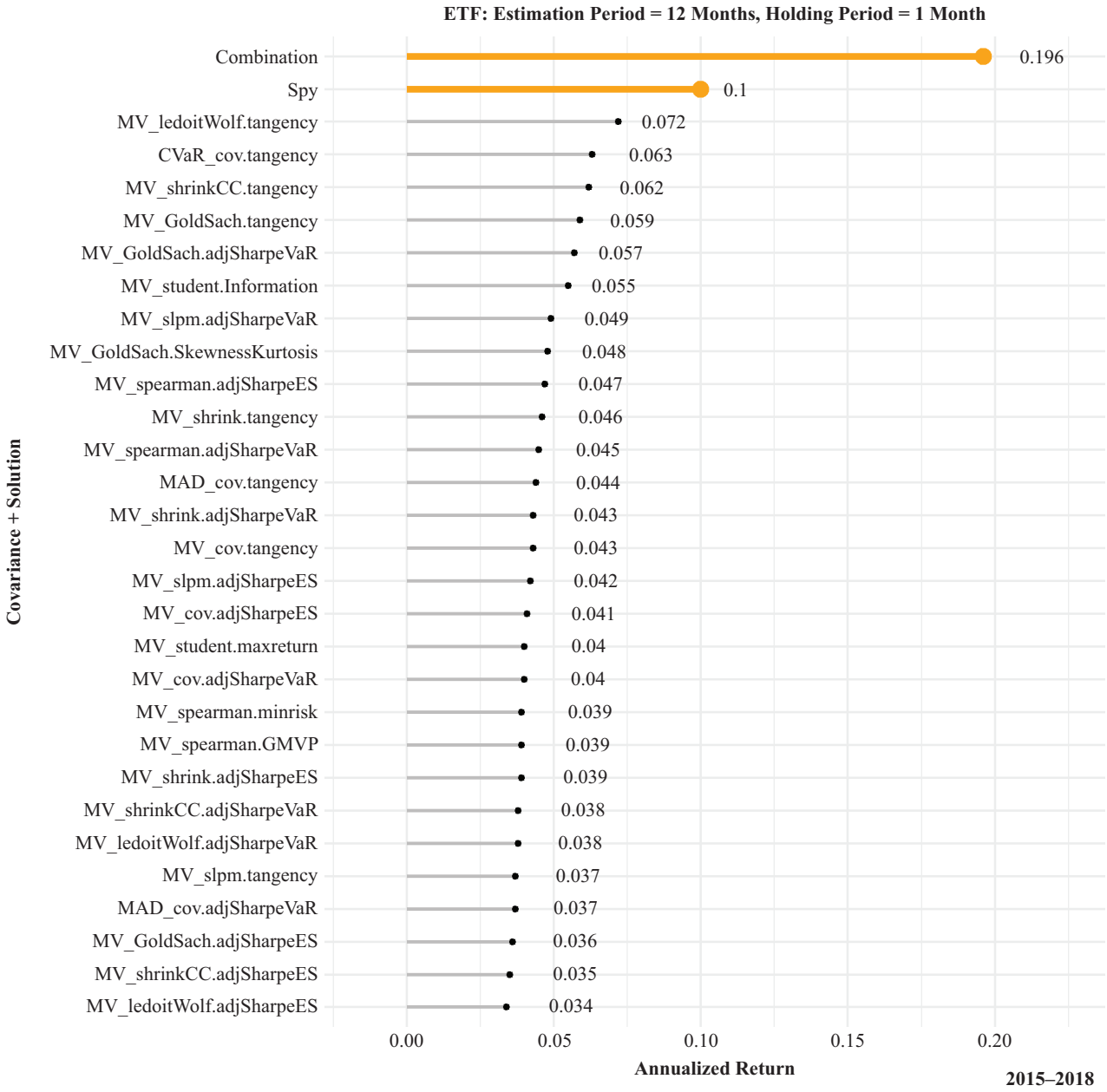
[4] We also measure the gap between the ex ante predictive returns and the ex post realized returns. This information is available to download as an online supplement at http://web.ntnu.edu.tw/_HMRD.pdf. This supplement indicates that the empirical optimality of a combination portfolio at T_1 is a good predictor for the optimality at T_2.

[5] This also allows us to avoid extreme multivariate correlation such as ±0.95.

Exhibit 11

Portfolios Ranked by Annualized Sharpe Ratio: ETF

Panel A: Annualized Return, Estimation = 12 Months, Holding = 1 Month

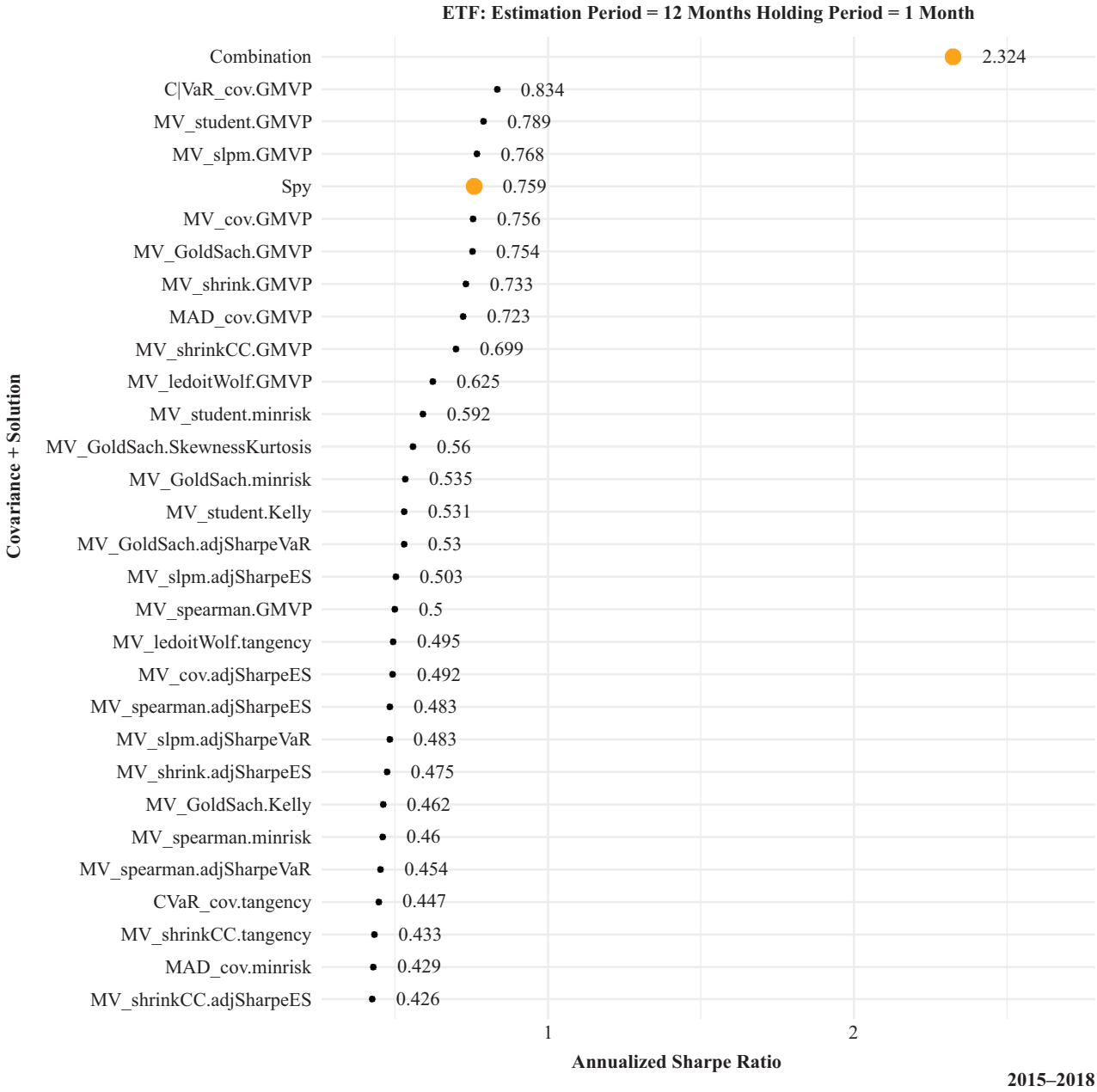


(continued)

EXHIBIT 11 *(continued)*

Portfolios Ranked by Annualized Sharpe Ratio: ETF

Panel B: Annualized Sharpe Ratio, Estimation = 12 Months, Holding = 1 Month



Notes: According to both panels, although the annualized expected returns of the combination portfolio and SPY are competitive (10% versus 19.6%), the annualized Sharpe ratio is quite significant (0.759 versus 2.324).

EXHIBIT 12

Comparison of Transaction Fee and Maximal Drawdown: ETF

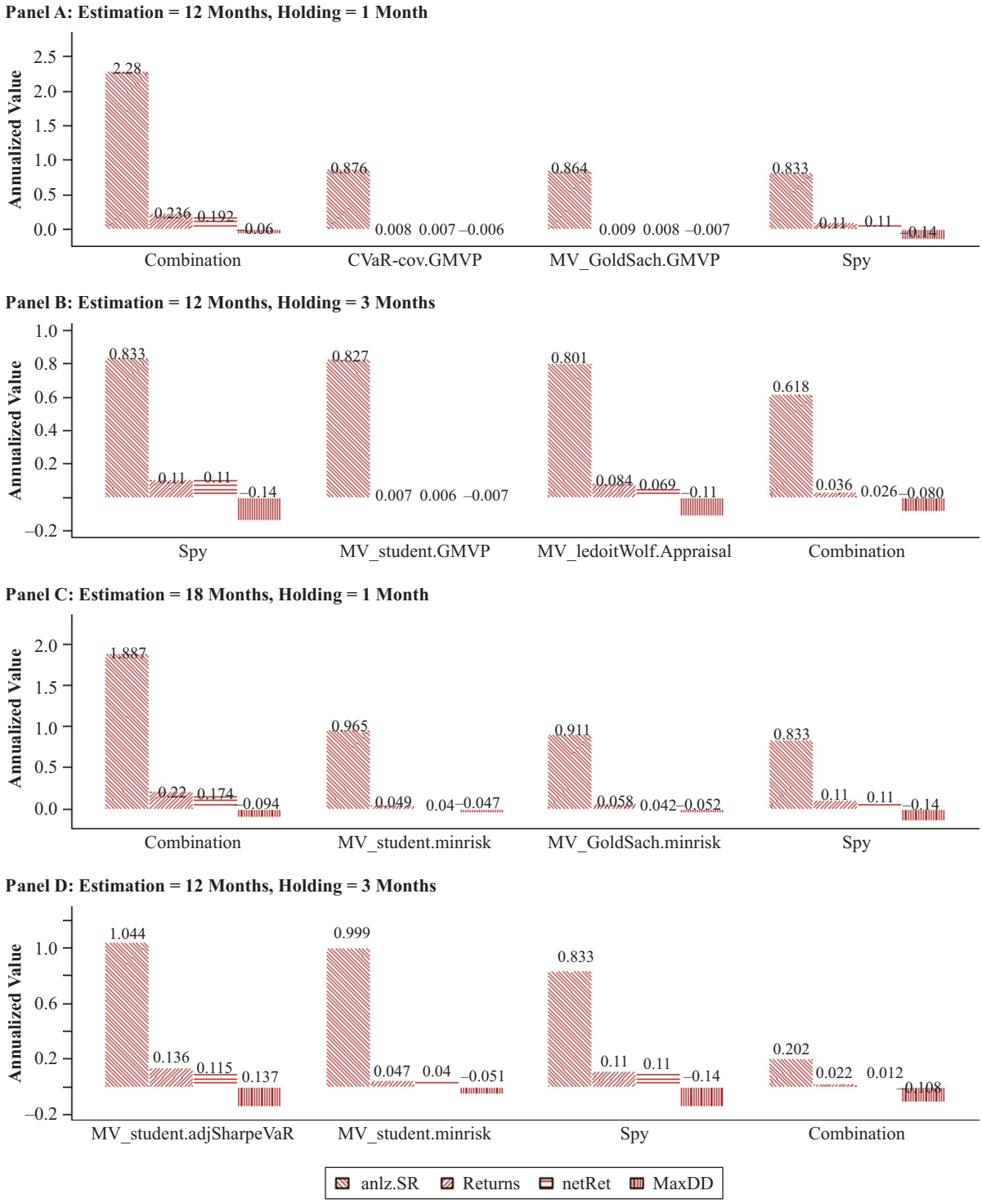
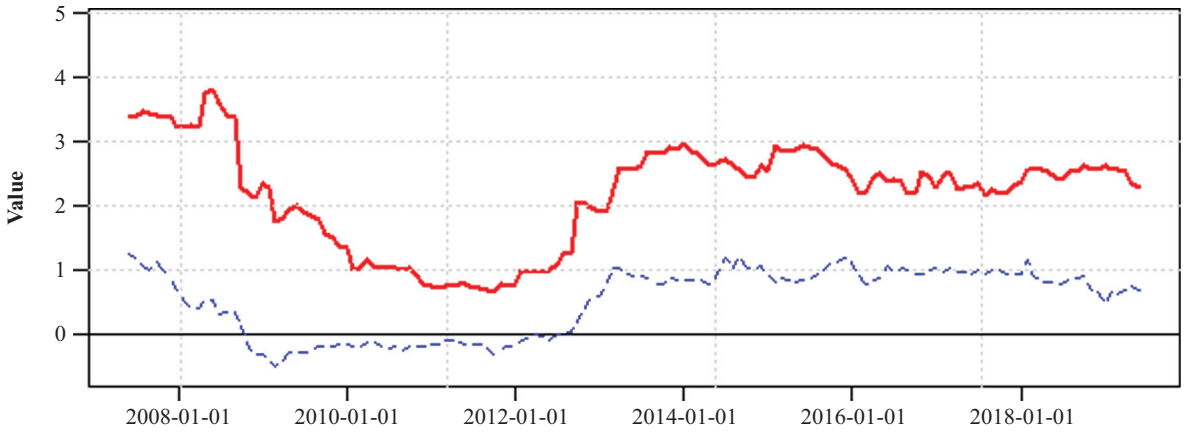
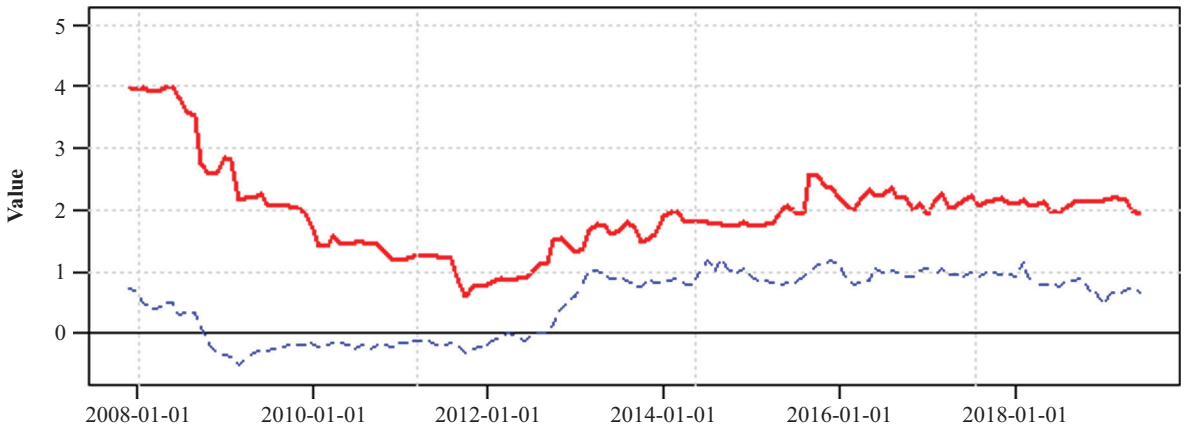


EXHIBIT 13
Comparison of Rolling Estimates: ETF

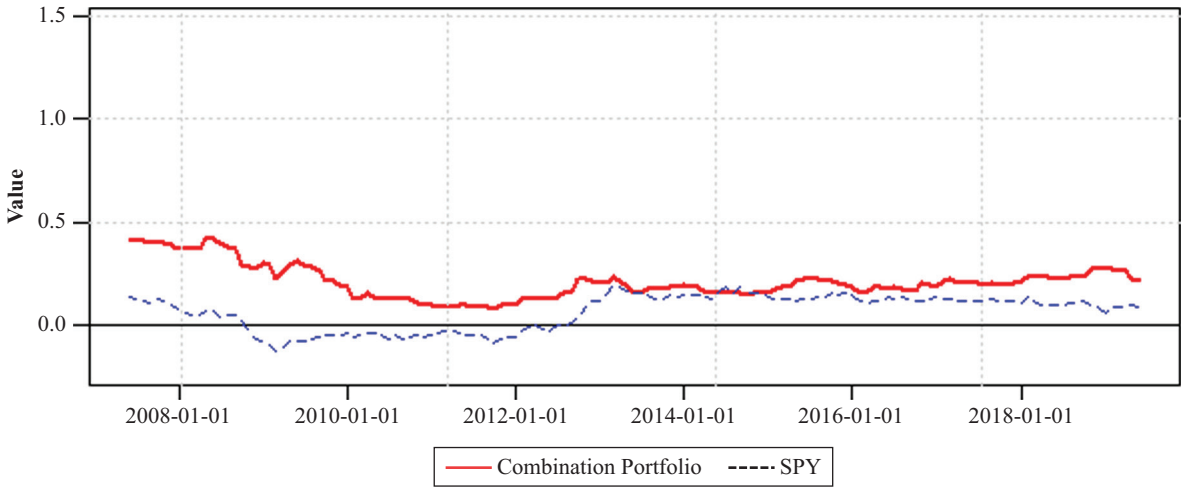
Panel A: Annualized Sharpe Ratio: Estimation = 12 Months, Holding = 1 Month



Panel B: Annualized Sharpe Ratio: Estimation = 18 Months, Holding = 1 Month



Panel C: Annualized Returns: Estimation = 12 Months, Holding = 1 Month



(continued)

EXHIBIT 13 *(continued)*

Comparison of Rolling Estimates: ETF

Panel D: Annualized Returns: Estimation = 18 Months, Holding = 1 Month

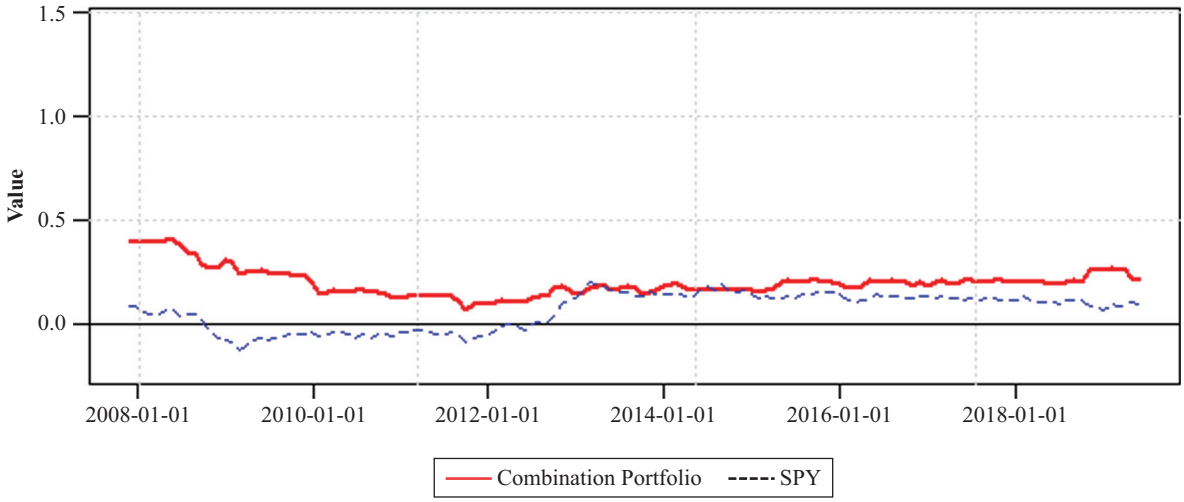


EXHIBIT 14

10,000 Simulation Results: Top 10 Portfolios Ranked by Annualized Sharpe Ratios

Rank	*N* = 10	Average anlz.SR	SD	Max.	Min.	Top 5	Top 10	Top 20
1	Combination	1.476	1.830	5.692	–3.706	2.9%	7.2%	31.7%
2	MV_student.maxreturn	1.232	2.105	6.047	–4.810	17.5%	24.1%	35.3%
3	MV_student.KellyRatio	1.219	2.042	5.847	–4.686	12.5%	19.6%	31.0%
4	MV_ledoitWolf.maxreturn	1.211	2.076	5.707	–4.578	4.0%	9.5%	22.2%
5	MV_shrinkCC.maxreturn	1.209	2.069	5.598	–4.544	3.1%	8.1%	21.0%
6	MV_ledoitWolf.KellyRatio	1.200	2.037	5.659	–4.581	2.2%	5.8%	15.9%
7	MV_shrink.maxreturn	1.200	2.051	5.402	–4.380	2.2%	7.5%	18.1%
8	MV_shrinkCC.KellyRatio	1.199	2.034	5.591	–4.567	1.7%	6.2%	15.0%
9	MV_shrink.KellyRatio	1.197	2.031	5.416	–4.533	1.5%	5.0%	15.3%
10	MAD_cov.maxreturn	1.196	2.045	5.474	–4.423	3.3%	7.2%	20.0%
Rank	***N* = 30**							
1	Combination	0.694	1.796	5.618	–3.082	4.1%	11.2%	52.8%
2	MV_student.tangency	0.077	2.077	6.532	–4.904	10.5%	13.7%	19.7%
3	MV_student.AppraisalRatio	0.075	2.043	6.445	–4.857	2.7%	6.4%	14.0%
4	MV_student.SharpeRatio.annualized	0.074	2.019	6.456	–4.865	2.8%	7.8%	14.6%
5	MV_student.BurkeRatio	0.073	1.979	6.409	–4.864	2.9%	6.1%	15.0%
6	MV_student.OmegaSharpeRatio	0.072	2.035	6.214	–4.821	5.4%	10.3%	15.8%
7	MV_student.BernardoLedoitRatio	0.072	2.035	6.214	–4.821	3.5%	9.5%	15.3%
8	MV_student.SortinoRatio	0.071	2.029	6.216	–4.827	3.6%	8.5%	15.2%
9	MV_student.SharpeRatio.annualizedES	0.071	1.996	6.309	–4.851	2.8%	6.6%	13.7%
10	MV_student.InformationRatio	0.070	1.871	6.096	–4.462	5.1%	8.6%	15.7%

Notes: N denotes the number of assets, and SD is the standard deviation of anlz.SR. Top 5 represents the percentage in which the corresponding portfolio is ranked in the top 5 out of 10,000; for example, 2.9% indicates that there are 290 counts in which the combination portfolio enters Top 5.

outperforms the competitor. The three right-hand-side columns measure the percentage that the corresponding portfolio ranked as the Top (5, 10, 20) out of 10,000 replications. The most important finding of the simulation is that as the number of assets N increases from 10 to 30, the combination portfolio's superior performance also increases.

CONCLUSION

In the machine learning literature, committees haven proven to be an effective way to improve predictive performance and robustness. To this end, this article develops a flexible framework to apply committees to portfolio selection. Backtesting results indicate that a large portfolio committee helps select the optimal portfolio. Moreover, the described framework is flexible and easy to implement and can use various types of portfolios as inputs, including maximal-drawdown, risk-budgeting, and risk-parity portfolios, among others. Although the portfolio committee is useful, more work is needed with regard to the problems of optimal timing and portfolio risk; moreover, the simulation described in the article may include more features such as various types of shocks. We leave these points for future work.

APPENDIX A

THE COMPUTATION OF SCORE

Like actual trading, portfolio selection is justified by out-of-sample performance, rather than in-sample asymptotic soundness. Our score is derived as follows:

Step 1. Compute 21 performance ratios for all assets in a 1-year period.
Step 2. At a given period, sorts all assets and choose the one with highest ratio.
Step 3. Repeat Step 2 for all ratios considered.
Step 4. Move 1 month ahead and iterate.

This four-step algorithm collects the rank map over time and across ratios, as illustrated by Exhibit A1. It indicates

EXHIBIT A1

Selected Performance Ratios' Pick for a 12-Month Estimation Window

	October 31, 1996	August 31, 1998	March 31, 2008	August 31, 2009	July 31, 2016
BernardoLedoitRatio	NKE	HD	KO	HD	MO
SkewnessKurtosisRatio	UTX	AAPL	JPM	HD	DWDP
CalmarRatio	NKE	PFE	KO	HD	MO
SterlingRatio	NKE	PFE	KO	HD	MO
BurkeRatio	NKE	HD	KO	HD	MO
KellyRatio	NKE	HD	KO	HD	MO
MartinRatio	NKE	HD	KO	HD	MO
PainRatio	NKE	HD	KO	HD	MO
SharpeRatio	NKE	HD	KO	HD	MO
SharpeRatioVaR	NKE	HD	KO	HD	MO
SharpeRatioES	NKE	PFE	KO	HD	MO
SharpeRatio.annualized	NKE	HD	KO	HD	MO
SharpeRatio.annualizedVaR	NKE	HD	KO	HD	MO
SharpeRatio.annualizedES	NKE	PFE	KO	HD	MO
AdjustedSharpeRatio	NKE	PFE	KO	HD	MO
AdjustedSharpeRatioVaR	UNH	BA	MO	MO	AXP
AdjustedSharpeRatioES	NKE	BA	MO	MO	AXP
AppraisalRatio	NKE	WMT	KO	HD	MO
InformationRatio	NKE	WMT	KO	MMM	MO
TreynorRatio	NKE	AAPL	KO	HD	MO
OmegaSharpeRatio	NKE	HD	KO	HD	MO
SortinoRatio	NKE	HD	KO	HD	MO

that, aside from both Sharpe ratios, other ratios usually arrive at a high level of consensus; namely, they usually select the same company. However, low variation (high consensus) will make our average score useless, and high variation (instability) will also make our average score useless. We find that six measures of both Sharpe ratios usually differ with stability; in fact, the six ratios never reach the same result. As a remedy, we use two Sharpe ratios adjusted by three risk measures as indicated by Equation 5. Although this method is data and not theory driven, we have found it useful for out-of-sample cross validation.

Appendix B

DESCRIPTIONS OF COVARIANCE MATRIX

Ledoit and Wolf (2003) proposed an estimator that shrinks the sample covariance toward the residual-based covariance of a single-index model. We label this covariance matrix *ledoitWolf*.

Ledoit and Wolf (2004) assumed a constant correlation as the shrinkage target and assumed that all pairwise correlations are identical, derived by averaging all pairwise sample correlations. This number together with the vector of sample variances implies the shrinkage target ***F***. We denote the matrix by *shrinkCC*.

Litterman and Winkelmann (1998) proposed a covariance estimator by weighting observations by the length of time. First, for variance, they defined the following formula:

$$\sigma_{ii}^{T}(M) = \frac{\sum_{s=0}^{T}(w_{T-s}r_{iT-s}^{2})}{\sum_{s=0}^{T}w_{T-s}}$$

where r_{it} denotes the returns on the i-th asset at date t, and w_t denotes the weight applied at date t. Litterman and Winkelmann (1998) assumed that the investment horizon is M days and that the variance of returns of asset i at time T over M is denoted by σ_{ii}^{t}, given zero mean returns. Accordingly, they defined the $\sigma_{ii}^{T}(M)$ as mentioned earlier. Second, they defined covariance as follows:

$$\sigma_{ij}^{T}(M) = \frac{\sum_{s=0}^{T}(w_{T-s}r_{iT-s}r_{jT-s})}{\sum_{s=0}^{T}w_{T-s}}$$

We choose a weight to be a declining function of time. For example, let today's weight be 1, yesterday's weight be 0.9, and the weight before yesterday be 0.81. Weighting data in this manner is also known as decaying data, and the decay rate is defined by Litterman and Winkelmann (1998) as follows:

$$1-\frac{w_{t-1}}{w_t}$$

In the present study, we set the decay rate as 0.1. We denote this covariance by *GoldSach*.

The fifth covariance matrix is the multivariate Student t distribution proposed by Kent, Tyler, and Vardi (1994), denoted by *student*, controlling for the adverse impacts of outliers. This estimator provides some degree of robustness to outliers without giving a high breakdown point. The random vector $\mathbf{Y}$ has a multivariate $t_{\nu}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ distribution if

$$\mathbf{Y} = \boldsymbol{\mu} + \sqrt{\frac{\nu}{\mathbf{W}}}\mathbf{Z}$$

where $\mathbf{W}$ denotes the χ^2 distribution with ν degrees of freedom, $\mathbf{Z}$ is $N_d(0, \boldsymbol{\Lambda})$ distributed, and $\mathbf{W}$ and $\mathbf{Z}$ are independent. The multivariate t distribution is a continuous scale mixture of multivariate normal distribution (Ruppert and Matteson 2015, pp. 165–167).

The sixth matrix is the multivariate Spearman rank–based covariance matrix, a nonparametric measure of covariance based on the ranks of variables. The Spearman measure is known to capture the dependencies when the random variables are not elliptically distributed. For simplicity, we denote it as *spearman*.

The seventh matrix is the symmetric lower partial moment, denoted by *slpm*. Like the semi-variance, lower partial moments (LPM) are asymmetric risk measures in that they consider information for only a portion of the return distribution. LPM are asymmetric because the computation of the lower and upper triangle of a covariance matrix uses different subsamples. Sometimes, LPM fails owing to the insufficient subsamples at one part. Therefore, we use the symmetric version.

The eighth and final matrix is the classic Bayesian shrinkage (Stein 1955), denoted by *shrink*.

ACKNOWLEDGMENTS

I appreciate two referees' insightful comments and the editorial assistance of Dr. Joseph Simonian, which substantially improved the style of writing. All errors are mine.

REFERENCES

Azzalini, A. 1985. "A Class of Distributions Which Includes the Normal Ones." *Scandinavian Journal of Statistics* 12 (2): 171–178.

Ben-Tal, A., L. El Ghaoui, and A. Nemirovski. 2009. *Robust Optimization*. Princeton, NJ: Princeton University Press.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Singapore: Springer-Verlag. 2009.

Candelon, B., C. Hurlin, and S. Tokpavi. 2012. "Sampling Error and Double Shrinkage Estimation of Minimum Variance Portfolios." *Journal of Empirical Finance* 19 (4): 511–527.

Chamberlain, G. 1983. "A Characterization of the Distributions That Imply Mean–Variance Utility Functions." *Journal of Economic Theory* 29 (1): 185–201.

Ellsberg, D. 1961. "Risk, Ambiguity, and the Savage Axioms." *Quarterly Journal of Economics* 75 (4): 643–669.

Fabozzi, F. J., P. N. Kolm, D. A. Pachamanova, and S. M. Focardi. 2007. "Robust Portfolio Optimization." *The Journal of Portfolio Management* 33 (3): 40–48.

Fama, E. 1965. "The Behavior of Stock Prices." *The Journal of Business* 38 (1): 34–105.

Garlappi, L., R. Uppal, and T. Wang. 2007. "Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach." *The Review of Financial Studies* 20 (1): 41–81.

Jagannathan, R., and T. Ma. 2003. "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps." *The Journal of Finance* 58 (4): 1651–1684.

Kent, J. T., D. E. Tyler, and Y. Vardi. 1994. "A Curious Likelihood Identity for the Multivariate *t* Distribution." *Communications in Statistics—Simulation and Computation* 23 (2): 441–453.

Kon, S. J. 1984. "Models of Stock Returns—A Comparison." *The Journal of Finance* 39 (1): 147–165.

Ledoit, O., and M. Wolf. 2003. "Improved Estimation of the Covariance Matrix of Returns with an Application to Portfolio Selection." *Journal of Empirical Finance* 10 (5): 603–621.

——. 2004. "Honey, I Shrunk the Sample Covariance Matrix." *The Journal of Portfolio Management* 30 (Summer): 110–119.

Litterman, R., and K. Winkelmann. 1998. "Estimating Covariance Matrices." *Risk Management Series*, Goldman Sachs.

Markowitz, H. M. 1952. "Portfolio Selection." *The Journal of Finance* 7 (1): 77–91.

Ruppert, D., and D. S. Matteson. *Statistics and Data Analysis for Financial Engineering—With R Examples*. New York: Springer Verlag. 2015.

Rustem, B., and M. Howe. *Algorithms for Worst-Case Design and Applications to Risk Management*. Princeton, NJ: Princeton University Press. 2002.

Simonian, J. 2011. "A Bayesian Approach to Building Robust Structural Credit Default Models." *Applied Economics Letters* 18 (14): 1397–1400.

——. 2019. "Portfolio Selection: A Game-Theoretic Approach." *The Journal of Portfolio Management* 45 (6): 108–116.

Simonian, J., and J. Davis. 2010. "Robust Value-at-Risk: An Information-Theoretic Approach." *Applied Economics Letters* 17 (16): 1551–1553.

——. 2011. "Incorporating Uncertainty into the Black–Litterman Portfolio Selection Model." *Applied Economics Letters* 18 (17): 1719–1722.

Stein, C. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pp. 197–206. Berkeley: University of California Press. 1955.

Tobin, J. 1958. "Liquidity Preference as Behavior Toward Risk." *The Review of Economic Studies* 25 (1): 65–85.

Uppal, R., and W. Tan. 2003. "Model Misspecification and Underdiversification." *The Journal of Finance* 58 (6): 2465–2486.

Weiqiang, J. G., K. G. Lim, M. Sim, and W. Zhang. 2012. "Portfolio Value-at-Risk Optimization for Asymmetrically Distributed Asset Returns." *European Journal of Operational Research* 221 (2): 397–406.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Portfolio Selection: *A Game-Theoretic Approach*

Joseph Simonian

The Journal of Portfolio Management
https://jpm.pm-research.com/content/45/6/108

ABSTRACT: *Game-theoretic methods are not widespread in finance. One reason is that practitioners do not see straightforward applications of game theory to core investment problems. To that end, in this article the author describes two novel portfolio construction applications of the game-theoretic solution concept known as the Shapley value. In the first, the Shapley value is used as a type of shrinkage operator to blend the allocation weights generated by two different portfolio selection frameworks. In the second, the Shapley value is used to build a unique approach to portfolio selection using what the author calls an optimization game. In both applications, asset classes are assumed to be the players who are bargaining over representation in a portfolio. As the article shows, the Shapley value naturally facilitates the construction of well-diversified portfolios that can satisfy specific investor goals without the need for additional formal constraints.*

Honey, I Shrunk the Sample Covariance Matrix

Olivier Ledoit and Michael Wolf

The Journal of Portfolio Management
https://jpm.pm-research.com/content/30/4/110

ABSTRACT: *The central message of this article is that no one should use the sample covariance matrix for portfolio optimization. It is subject to estimation error of the kind most likely to perturb a mean–variance optimizer. Instead, a matrix can be obtained from the sample covariance matrix through a transformation called shrinkage. This tends to pull the most extreme coefficients toward more central values, systematically reducing estimation error when it matters most. Statistically, the challenge is to know the optimal shrinkage intensity. Shrinkage reduces portfolio tracking error relative to a benchmark index and substantially raises the manager's realized information ratio.*

A Modified Hierarchical Risk Parity Framework for Portfolio Management

MARAT MOLYBOGA

MARAT MOLYBOGA is the director of research at Efficient Capital Management in Warrenville, IL.
molyboga@efficient.com

KEY FINDINGS

- The author introduces a modified hierarchical risk parity (MHRP) approach, which incorporates three popular portfolio management techniques into the hierarchical risk parity framework of López de Prado.
- He reports a striking improvement in the out-of-sample Sharpe ratio of 50%, on average, for portfolios of commodity trading advisors.
- The author argues that the MHRP framework has broad applications for portfolios of traditional and alternative investments.

ABSTRACT: *This article introduces a modified hierarchical risk parity (MHRP) approach that extends the HRP approach by incorporating three intuitive elements commonly used by practitioners. The new approach (1) replaces the sample covariance matrix with an exponentially weighted covariance matrix with Ledoit–Wolf shrinkage; (2) improves diversification across portfolio constituents both within and across clusters by relying on an equal volatility, rather than an inverse variance, allocation approach; and (3) improves diversification across time by applying volatility targeting to portfolios. The author examines the impact of the enhancements on portfolios of commodity trading advisors within a large-scale Monte Carlo simulation framework that accounts for the realistic constraints of institutional investors. The author finds a striking improvement in the out-of-sample Sharpe ratio of 50%, on average, along with a reduction in downside risk.*

TOPICS: ***Statistical methods, simulations, big data/machine learning****

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

Portfolio selection has been a fruitful area of research since the introduction of the parsimonious theory of Markowitz (1952), which reduces a complex asset allocation problem to a simple calculation that relies solely on the vector of expected returns and the covariance matrix. Unfortunately, as discussed by Michaud (1989) and Best and Grauer (1991), this beautiful theory produces ugly results, as evidenced by poor out-of-sample performance and extreme, unstable portfolio weights. The issue of instability is driven by the high sensitivity of the Markowitz portfolio weights to estimation error.[1] This difficulty is further

[1] Markowitz portfolios are obtained by solving a quadratic problem, which requires the inversion of a covariance matrix. Bailey and López de Prado (2012) and Yang et al. (2016) showed that the magnitude of the sensitivity issue can be assessed using the condition number of the covariance matrix. If the covariance matrix is near singular, the condition number, defined as the ratio of the largest to the smallest eigenvalues of

compounded by the impossible problem of estimating the vector of expected returns with a high degree of precision, as noted by Merton (1980).

Although many articles have attempted to reduce the impact of estimation error, DeMiquel, Garlappi, and Uppal (2009) showed that even very sophisticated approaches[2] fail to outperform a naive 1/N portfolio and argued that the estimation window required to capture the potential gains of optimal portfolios is too long. The failure of mean–variance optimization has led to the emergence of the popular risk-parity approach discussed in detail by Maillard, Roncalli, and Teiletche (2010). The risk-parity approach ignores return forecasts and instead attempts to maximize diversification by allocating risk equally across portfolio constituents. Despite its popularity, the risk parity approach is subject to the same weaknesses as mean–variance optimization, namely, unstable weights caused by an inversion of a near singular covariance matrix and a lack of alignment with the top-down perspective in portfolio management typically employed by sophisticated practitioners. The hierarchical risk parity (HRP) approach introduced by López de Prado (2016) elegantly addresses those two weaknesses, establishing it as the cutting-edge application of machine learning to portfolio management.

The purpose of this article is to examine whether the HRP approach can be further improved by adding three intuitive elements commonly used by practitioners. The modified approach, MHRP, (1) replaces the sample covariance matrix with an exponentially weighted covariance matrix with Ledoit–Wolf shrinkage; (2) improves diversification across portfolio constituents both within and across clusters by relying on an equal volatility approach rather than an inverse variance approach; and (3) improves diversification across the dimension of time by applying volatility targeting to portfolios.

Although the proposed improvements represent relatively minor changes to HRP from a theoretical perspective, the practical implications for investors can be meaningful because the individual components have been tested across many asset classes and market cycles and subsequently embraced by practitioners for their robustness and solid theoretical underpinnings. For example, exponential weighting takes into account the presence of heteroskedasticity in financial returns, which is at the heart of RiskMetrics' volatility estimation, as discussed by JPMorgan and Reuters (1996), while also being closely related to the illustrious generalized autoregressive conditional heteroskedasticity (GARCH) model of Bollerslev (1986). Similarly, diversification has been proclaimed by the Nobel Prize winner Harry Markowitz to be "the only free lunch in finance." Improvements in performance as a result of greater diversification across portfolio constituents with equal risk allocation and across time with portfolio volatility targeting have been reported in many previous studies, such as those by Hallerbach (2012), Kim, Tse, and Wald (2016), Daniel and Moskowitz (2016), and Moreira and Muir (2017).

I examine the marginal impact of each component and their combined impact on the performance of portfolios of commodity trading advisors (CTAs) within the large-scale simulation framework of Molyboga and L'Ahelec (2016) that accounts for the realistic constraints faced by institutional investors. The framework is imposed on a BarclayHedge CTA sample of 528 live and 1,113 defunct funds over the 2002–2016 period. I find that each enhancement improves out-of-sample Sharpe ratios of multi-CTA portfolios by 13% to 19%, on average. Moreover, when all three enhancements are combined into a unified MHRP approach, I observe a striking improvement in the out-of-sample Sharpe ratio of 50%, on average, with a meaningful reduction in downside risk.

Finally, I discuss whether CTA investors can capture the benefit of MHRP after accounting for real-life constraints such as trading frictions caused by monthly rebalancing and portfolio leverage changes. I show that although the frictions are potentially high for fund investors because of a time gap between a redemption from one fund and a subscription into another fund, the frictions can be mitigated if the investor uses managed accounts that employ notional funding and benefit from more favorable liquidity terms. Although the performance benefits of the MHRP approach should be carefully evaluated within other types of portfolios that have their own sets of investments and are subject to their own unique real-life constraints, the three

the covariance matrix, is large and the portfolio weights are highly sensitive to estimation error.

[2]DeMiquel, Garlappi, and Uppal (2009) considered 14 models that include the sample-based mean–variance model and its extensions designed to reduce estimation error. Two of the models they considered are the Pastor (2000) model, which produces Bayesian portfolios based on belief in an asset-pricing model, and the MacKinlay and Pastor (2000) approach, which produces portfolios implied by asset-pricing models with unobservable factors.

proposed enhancements are likely to have a positive impact on the portfolio performance of both traditional and alternative investments because they improve the quality of the covariance matrix estimates and the degree of diversification across portfolio constituents and time. The rest of the article is organized as follows. The first section describes the HRP approach and the three enhancements that comprise the MHRP approach. The second section describes the data and the empirical framework of Molyboga and L'Ahelec (2016). The third section presents empirical results, and the fourth section concludes.

FROM HRP TO MHRP

As discussed by López de Prado (2016), most bottom-up approaches to portfolio construction, such as mean–variance optimization or risk parity, tend to rely heavily on each value of a correlation matrix, implying that any two securities are potential substitutes for one another. This assumption is inconsistent with the investment practices of investment professionals, particularly in light of the work of Brinson, Hood, and Beebower (1995), who highlighted the importance of asset allocation decisions. If an investor attempts to build a 60–40 portfolio of stocks and bonds, which is a typical starting point for a US institutional investor according to Anson (2011), the investor would group stocks and bonds separately rather than consider individual stocks and bonds as substitutes for each other. López de Prado (2016) showed that considering a hierarchical tree structure allows for focusing on a small number of relationships that are consistent with the top-down perspective of investors who build their portfolios by starting at the asset-class level and then going down to the level of the individual securities.

As shown by López de Prado (2016), the HRP approach follows three steps[3]:

1. **Tree clustering.** Hierarchical clustering is performed using a sample correlation matrix. This step highlights important top-down relationships among individual portfolio constituents and their clusters.
2. **Quasi-diagonalization.** Similar investments are placed together and dissimilar investments are placed far apart. This procedure reshuffles rows and columns of the original correlation matrix so that the largest values are close to the diagonal and the smallest values are away from the diagonal.
3. **Recursive bisection.** This step allocates across and within clusters using an inverse variance allocation (IVA) approach.

In this article, I enhance the HRP approach by adding three intuitive elements commonly used by practitioners. This modified approach (1) replaces the sample covariance matrix with an exponentially weighted covariance matrix with Ledoit–Wolf shrinkage, (2) improves diversification across portfolio constituents both within and across clusters by relying on an equal volatility allocation (EVA) approach rather than an IVA approach, and (3) improves diversification across the dimension of time by applying volatility targeting to portfolios.

It is important to distinguish the volatility weighting of portfolio constituents (which closely relates to the concept of diversification across portfolio constituents) from targeting portfolio volatility (which is linked to diversification across time). Volatility weighting is a constant leverage approach highlighted by Maillard, Roncalli, and Teiletche (2010) and Hallerbach (2012) that allocates risk equally across portfolio constituents. By contrast, volatility targeting dynamically scales aggregate portfolio leverage to achieve constant expected portfolio volatility or equally allocates risk across time rather than among portfolio constituents.

Covariance Matrix Estimation

Exponential weighting is used pervasively in academia and industry to account for the presence of heteroskedasticity in financial returns. The exponentially weighted moving-average approach is at the heart of the volatility estimation employed by RiskMetrics, as discussed by JPMorgan and Reuters (1996), and is closely related to the GARCH model of Bollerslev (1986). Likewise, Pafka, Potters, and Kondor (2004) extensively examined applications of exponentially weighted covariance matrices to portfolio optimization problems.

One of the most common approaches used to improve the estimation of covariance matrixes is based on the shrinkage approach of Ledoit and Wolf (2004).

[3] See López de Prado (2016) for a detailed description of the HRP approach and the Python implementation.

I follow Molyboga (2019) and combine shrinkage estimation with exponential weighting of the covariance matrix.

Better Diversification across Portfolio Constituents

This study considers an IVA approach and an EVA approach. The two approaches can be contrasted as follows:

- IVA is a methodology in which an allocation weight to portfolio constituent i is inversely related to the estimated variance of the constituent, $\hat{\sigma}_i^2$, as follows:

$$w_i^{IVA} = \frac{1/\hat{\sigma}_i^2}{\sum_{j=1}^{N} 1/\hat{\sigma}_j^2}.$$

- EVA is an equal volatility (or inverse volatility) allocation wherein

$$w_i^{EVA} = \frac{1/\hat{\sigma}_i}{\sum_{j=1}^{N} 1/\hat{\sigma}_j}.$$

López de Prado (2016) suggested that IVA is optimal when the covariance matrix is diagonal because it produces the minimum variance portfolio. Because investors care about risk-adjusted performance measures such as the Sharpe ratio, IVA is likely suboptimal because it only focuses on minimizing risk rather than on optimizing the trade-off between return and risk. Section A in the online appendix considers two potential assumptions regarding portfolio constituents. It shows that if all portfolio constituents have the same expected return, minimization of variance is equivalent to Sharpe maximization and IVA maximizes the Sharpe ratio. However, if the portfolio constituents have similar Sharpe ratios, which is more likely than similar expected returns, EVA is optimal.

Practitioners often prefer EVA because it attempts to maximize diversification by equally allocating risk, measured in terms of volatility, across portfolio constituents, and it intuitively responds to changes in the volatility levels of portfolio constituents.[4] To demonstrate the utility of this approach, let us consider a simple hypothetical example in which fund A exhibits annualized volatility of 20% and fund B realizes annualized volatility of 10%. Because fund A is twice as risky as fund B, an investor who attempts to diversify risk equally across the two funds would allocate one-third of the investment amount to fund A and two-thirds to fund B ($1/3 \times 20\% = 2/3 \times 10\%$). By contrast, the IVA approach would give a one-fifth allocation to fund A and a four-fifths allocation to fund B—an allocation of twice as much risk to fund B as to fund A ($2 \times [1/5 \times 20\%] = 4/5 \times 10\%$).

CTAs commonly offer their trading programs at multiple volatility levels, capitalizing on the free leverage embedded in futures and forward contracts that are traded on margin. If manager A decides to offer a 10% version of the program, both funds run at the same volatility and both the EVA and IVA approaches will allocate risk equally between the funds. Thus, EVA is more attractive to practitioners because it preserves risk allocations across funds regardless of the choice of volatility levels while attempting to maximize diversification by allocating risk equally across portfolio constituents.

Better Diversification across the Dimension of Time

Previous research has suggested that managing portfolio volatility improves the performance of risk premiums and active strategies. Barroso and Santa-Clara (2015) and Daniel and Moskowitz (2016) showed that volatility targeting, or adjusting exposure to target a constant ex ante volatility, nearly doubles the Sharpe ratio of cross-sectional momentum. Kim, Tse, and Wald (2016) reported that the abnormal returns of time-series momentum are largely driven by scaling returns to volatility. Moreira and Muir (2017) found evidence for the benefit of managing the volatility of portfolios across a broad array of risk premiums, such as market, value, currency carry, and betting against beta.

Molyboga (2019) provided theoretical underpinnings for volatility targeting and investigated empirically whether, and under what conditions, it improves the performance of multi-CTA portfolios. Molyboga (2019) reported that volatility targeting improves the out-of-sample, net-of-fee Sharpe ratio of portfolios with 10 CTAs by approximately 13%–14%, on average. This improvement in the Sharpe ratio represents approximately 63–68

[4] As shown in Section A of the online appendix, this approach is optimal when portfolio constituents have equal Sharpe ratios.

bps of return per annum for a 15% volatility portfolio. As the number of managers in the portfolio varies between 5 and 20, volatility targeting consistently improves performance by 0.53% to 0.80% per annum, with larger portfolios achieving greater performance enhancements. Section B in the online appendix describes how volatility targeting is accomplished through dynamically adjusting portfolio leverage.

Portfolio Construction Approaches

To examine the marginal impact of each enhancement and their combined impact on performance, I consider the following four portfolio construction approaches:

- *HRP* is the original hierarchical risk parity approach of López de Prado (2016).
- HRP_C is an HRP approach that relies on the exponentially weighted covariance matrix with Ledoit–Wolf shrinkage.
- HRP_{CE} is an HRP approach that relies on the exponentially weighted covariance matrix with Ledoit–Wolf shrinkage and uses the EVA instead of the IVA to diversify within and across clusters.
- *MHRP* is a modified HRP approach that relies on the exponentially weighted covariance matrix with Ledoit–Wolf shrinkage, uses the EVA instead of the IVA to diversify within and across clusters, and targets a constant portfolio volatility of 12%.

The marginal impact of the covariance matrix estimation is evaluated by comparing the performance of HRP_C against HRP. The marginal improvement due to superior diversification across and within clusters is measured by comparing the performance of HRP_{CE} against that of HRP_C. The marginal contribution of volatility targeting is measured by comparing the performance of MHRP against HRP_{CE}. Finally, the combined impact of the three enhancements is captured by comparing the performance of MHRP against HRP.

DATA AND EMPIRICAL FRAMEWORK

In this section, I describe the data used in the study and the empirical framework of Molyboga and L'Ahelec (2016) used to evaluate the portfolio management approaches.

Data

I rely on the BarclayHedge database, the largest and highest-quality publicly available database of CTAs, according to Joenvaara, Kosowski, and Tolonen (2012). I include the graveyard database of defunct funds to account for survivorship bias and adjust for backfill/incubation bias by using the first reported date, as recommended by Bhardwaj, Gorton, and Rowenhorst (2014), Foran et al. (2017), and Jorion and Schwarz (2017).[5] Bhardwaj, Gorton, and Rowenhorst (2014) showed that the common adjustment of removing 24 months of returns is inadequate for CTAs because the remaining bias is still more than 1% per annum in a value-weighted index and almost 3% in an equally weighted index. Because the first reported date in the BarclayHedge database is December 2002, our dataset is limited to December 2002 through December 2017. Although the selection bias that results from funds choosing not to report to public databases is the most difficult to avoid, Edelman, Fung, and Hsieh (2013) demonstrated that it is likely to be insignificant.[6] I use the liquidation estimate of 1% as suggested by Ackermann, McEnally, and Ravenscraft (1999). Because CTA investments are often chosen for their diversification benefit during crisis periods, as summarized by Molyboga and L'Ahelec (2016), I exclude fund managers that are most susceptible to market crisis, such as those that predominantly rely on options or arbitrage.[7]

The final bias-free dataset includes 576 active and 1,187 defunct funds with net returns between December

[5] The voluntary nature of self-reporting to hedge fund databases results in the backfill/incubation bias. Before reporting to public databases, funds often go through an incubation period during which they build a track record using proprietary capital. If the track record is attractive, fund managers report their performance to a database and often backfill the returns generated before inclusion in the database. This practice results in an upward bias because funds with poor performance are unlikely to report their performance to the database.

[6] There are two types of funds that choose not to report in public databases: poorly performing funds that cannot attract new investors based on their track record and successful funds that do not rely on the commercial databases to raise assets. Edelman, Fung, and Hsieh (2013) showed that the two sources of nonreporting bias cancel each other out.

[7] I use the BarclayHedge classification field to remove funds with three classifications: option strategies, stock index/option strategies, and arbitrage. I repeat analysis after including the three classifications and find qualitatively similar results.

EXHIBIT 1
Summary Statistics of CTAs for 2002–2017

	Defunct	Live	All
Mean	0.05%	0.32%	0.14%
Median	–0.02%	0.27%	0.08%
St Dev	3.69%	3.85%	3.74%
Skewness	0.12	0.17	0.14
Excess kurtosis	1.09	1.10	1.09
Autocorrelation	–0.01	–0.07	–0.03
Average no. of managers	325	226	551
Average track record length	50	71	57
Average AUM	164	819	379
Total no. of managers	1,187	576	1,763
No. of observations	58,829	40,978	99,807

2002 and December 2017. I use monthly values of the three-month T-bill secondary market rate (ticker: TB3MS) from FRED as a proxy for the risk-free rate.

Exhibit 1 reports the average summary statistics for the defunct managers, live managers, and all managers in the CTA sample. The statistics include mean monthly excess return, median monthly excess return, monthly standard deviation, skewness, excess kurtosis, autocorrelation of order one, average number of managers, average track record length, average assets under management (AUM) in millions of US dollars, total number of managers, and number of observations.

Consistent with the findings of previous studies, live managers tend to perform better, stay in business longer, and attract more assets than defunct managers. The mean monthly excess return of live managers is 0.32%, which at the 99% significance level exceeds the 0.05% mean monthly excess return of the defunct managers. The average track record length of live and defunct managers is 71 and 57 months, respectively. The average AUM of live and defunct managers is US$819 million and US$164 million, respectively. The high mortality rate among CTA funds is indicated by the large share of defunct funds in the database. Out of 1,763 funds in the sample, 1,187 funds are defunct and only 576 are live.

Empirical Analysis

In this section, I describe the large-scale simulation framework of Molyboga and L'Ahelec (2016). As discussed in the previous section, this article attempts to improve on the original HRP approach of López de Prado (2016) by adding three intuitive enhancements commonly used by practitioners. The MHRP approach (1) replaces a sample covariance matrix with an exponentially weighted covariance matrix with Ledoit–Wolf shrinkage, (2) improves diversification across portfolio constituents within and across clusters by relying on an EVA rather than an IVA approach, and (3) improves diversification across the dimension of time by volatility targeting portfolios. To examine the marginal impact of each enhancement and their combined impact on performance, I consider the four portfolio construction approaches: HRP, HRP_C, HRP_{CE}, and MHRP.

In this article, I apply the large-scale framework of Molyboga and L'Ahelec (2016) with 2,000 simulations for the out-of-sample period between January 2006 and December 2017. The simulation framework is designed to evaluate portfolio construction approaches subject to the realistic constraints of institutional investors. The first allocation decision is made at the end of December 2005. Because of the delay in CTA reporting documented by Molyboga, Baek, and Bilson (2017), an institutional investor can rely on information only through November 2005 for the January 2006 allocation. The simulation accounts for two standard constraints that institutional investors face by considering only funds with at least 36 months of returns and excluding all funds in the bottom quintile of AUM among the funds considered. The simulation then randomly selects 15 funds from the available pool of funds and applies the four portfolio construction approaches. For each portfolio approach, monthly returns are recorded for January 2006 using the liquidation bias adjustment, if needed. At the end of January 2006, the available fund pool is updated to comply with the previously noted requirements of institutional investors. All funds from the original portfolio that no longer meet the requirements are randomly replaced with funds from the new pool. The portfolio turnover is relatively low because the funds that continue meeting the selection criteria remain in the portfolio. Each portfolio is rebalanced using one of the four methodologies, and the process is repeated for the entire out-of-sample period covering January 2006 through December 2017. A single simulation run produces four time series of out-of-sample performance, one for each portfolio construction approach. Distributions of out-of-sample returns are constructed based on the results of 2,000 simulations.

EXHIBIT 2
Annual Statistics of CTAs

Date	AUM	CTAs
2006	11.88	148
2007	10.54	156
2008	12.21	171
2009	12.78	200
2010	14.82	222
2011	14.02	244
2012	12.20	249
2013	10.87	242
2014	10.45	223
2015	10.01	213
2016	9.67	215
2017	10.21	223

Exhibit 2 reports the average AUM threshold level for each year (in millions of US dollars) and the average number of active funds that have at least 36 months of returns and meet the AUM threshold. Their time-series variation reflects the dynamic nature of the CTA universe.

EMPIRICAL RESULTS

In this section, I present empirical out-of-sample results obtained within the large-scale simulation framework. I start by measuring the performance impact of improved covariance matrix estimation, diversification across portfolio constituents, and volatility targeting. At this stage, I largely focus on the average impact on out-of-sample performance, but I also consider the quartiles of the distributions of Sharpe ratios. I then evaluate the combined improvement of the enhancements by comparing the MHRP and the HRP approaches. Finally, I discuss implications for CTA investors as well as broad applications to other types of traditional and alternative investments.

Marginal Impact of the New Covariance Matrix Estimation Approach

The large-scale simulation framework produces distributions of out-of-sample performance. Molyboga, Baek, and Bilson (2017) introduced a methodology for comparing distributions that relies on utility functions and stochastic dominance. In this article, I primarily focus on means and medians but also consider distribution quantiles based on this methodology. I start the analysis by comparing the performance of the original HRP approach that uses the sample correlation matrix for clustering and calculating volatilities with the HRP_C approach that relies on the correlation matrix estimated with exponential weighting and Ledoit and Wolf (2004) shrinkage.

EXHIBIT 3
Statistics of the Distributions of Sharpe Ratios for HRP and HRP_C

	HRP	HRP_C
Mean	0.217	0.248*
St Dev	0.255	0.227
Max	1.155	1.168
Third Quartile	0.386	0.403
Median	0.216	0.247
First Quartile	0.045	0.094
Min	–0.641	–0.474

*Notes: *Significant at the 1% level.*

Exhibit 3 reports the mean, standard deviation, and quantiles of the distributions of Sharpe ratios, including the minimum, first quartile, median, third quartile, and maximum values. The distributions of Sharpe ratios are generated for the two portfolio construction approaches of HRP and HRP_C within the large-scale simulation framework. The out-of-sample period covers January 2006 to December 2017. The P-values of the differences in the means of the Sharpe ratios are calculated using the bootstrap approach of Molyboga and L'Ahelec (2016).

The improvement due to the superior covariance matrix estimation is meaningful in both economic and statistical terms. The four quartiles of the HRP_C method are consistently higher than those of the original HRP allocation of López de Prado (2016), and the standard deviation is lower. The mean Sharpe ratio of 0.248 for the HRP_C approach is approximately 14.2% higher than the mean Sharpe ratio of 0.217 for the standard HRP approach. This difference is significant at the 99% confidence level[8] and highlights the importance of de-noising correlation matrixes. Although the suggested covariance

[8] Statistical inference uses the bootstrap methodology of Molyboga and L'Ahelec (2016). Because the simulations are not independent, a bootstrap approach is required to eliminate the bias inherent in standard tests, as discussed in detail by Molyboga, Baek, and Bilson (2017).

EXHIBIT 4

Statistics of the Distributions of Sharpe Ratios for HRP_C and HRP_{CE}

	HRP_C	HRP_{CE}
Mean	0.248	0.294*
St Dev	0.227	0.146
Max	1.168	0.853
Third Quartile	0.403	0.390
Median	0.247	0.291
First Quartile	0.094	0.199
Min	–0.474	–0.203

*Notes: *Significance at the 1% level.*

matrix estimation technique that relies on exponential weighting and Ledoit and Wolf (2004) shrinkage improves the out-of-sample performance of the original HRP approach, future research can extend these findings by investigating whether alternative approaches to de-noising the correlation matrixes, such as the random matrix theory approach of Pafka, Potters, and Kondor (2004), can further improve performance.

Marginal Impact of Improved Diversification across Portfolio Constituents

Having established the benefit of employing an exponentially weighted covariance matrix with shrinkage estimation, I turn to the performance impact of portfolio diversification across portfolio constituents. As discussed earlier, EVA is more attractive to practitioners than IVA because it preserves risk allocations across funds regardless of their choice of volatility levels and attempts to maximize diversification by allocating risk equally across portfolio constituents.

Exhibit 4 reports the mean, standard deviation, and quantiles of the distribution of Sharpe ratios, including the minimum, first quartile, median, third quartile, and maximum values. The distributions of Sharpe ratios are generated for the two portfolio construction approaches of HRP_C and HRP_{CE} within the large-scale simulation framework.

The improvement due to the superior diversification across portfolio constituents is also strong, and the standard deviation of the distribution decreases by about one-third; however, the improvement is less consistent across the distribution of Sharpe ratios. The performance improves for the bottom three quartiles, with a slight degradation for the top quartile. The mean Sharpe ratio of 0.294 for the HRP_{CE} approach that relies on EVA across and within clusters is approximately 18.6% higher than the mean Sharpe ratio of 0.248 for the HRP_C approach that uses IVA. This difference is significant at the 99% confidence level and suggests that an increase in portfolio diversification within the HRP approach leads to superior out-of-sample performance.

Marginal Impact of Improved Diversification across Time

Finally, I examine the impact of improving diversification across time with volatility targeting. As a reminder, it is important to distinguish volatility weighting of portfolio constituents, which closely relates to the concept of diversification across portfolio constituents, from targeting portfolio volatility, which is linked to diversification across time. Although EVA distributes risk equally across portfolio constituents while maintaining constant leverage at the portfolio level, volatility targeting dynamically scales aggregate portfolio leverage to achieve constant expected portfolio volatility by allocating risk equally across time rather than portfolio constituents (see section B in the online appendix for details).

Exhibit 5 reports the mean, standard deviation, and quantiles of the distribution of Sharpe ratios, including the minimum, first quartile, median, third quartile, and maximum values. The distributions of Sharpe ratios are generated for the two portfolio construction approaches of HRP_{CE} and MHRP within the large-scale simulation framework.

The improvement due to the superior diversification across time is consistent across the distribution of Sharpe ratios, except for a few outliers in the left tail. The mean Sharpe ratio of 0.333 for the MHRP approach that relies on volatility targeting to diversify across time is approximately 13.1% higher than the mean Sharpe ratio of 0.294 for the HRP_{CE} approach. This difference is significant at the 99% confidence level and suggests that volatility targeting improves performance. This finding is consistent with those of Kim, Tse, and Wald (2016) and Moreira and Muir (2017), who reported a positive impact from volatility targeting.

EXHIBIT 5

Statistics of the Distributions of Sharpe Ratios for HRP_{CE} and MHRP

	HRP_{CE}	MHRP
Mean	0.294	0.333
St Dev	0.146	0.149
Max	0.853	0.916
Third Quartile	0.390	0.431
Median	0.291	0.336
First Quartile	0.199	0.234
Min	–0.203	–0.211

EXHIBIT 6

Statistics of the Distributions of Sharpe Ratios for HRP and MHRP

	HRP	MHRP
Mean	0.217	0.333*
St Dev	0.255	0.149
Max	1.155	0.916
Third Quartile	0.386	0.431
Median	0.216	0.336
First Quartile	0.045	0.234
Min	–0.641	–0.211

*Notes: *Significance at the 1% level.*

Combined Impact of the Three Enhancements: MHRP versus HRP

Having established that each of the three enhancements improves out-of-sample performance by 13% to 19%, on average, I examine their combined impact on performance by comparing the MHRP approach to the original HRP portfolio of López de Prado (2016).

Exhibit 6 reports the mean, standard deviation, and quantiles of the distribution of Sharpe ratios, including the minimum, first quartile, median, third quartile, and maximum values. The distributions of Sharpe ratios are generated for the two portfolio construction approaches of HRP and MHRP within the large-scale simulation framework. The improvement is striking, with the mean Sharpe ratio increasing by more than 50% and the standard deviation decreasing by approximately 40%.

Distributions of Sharpe ratios can be visualized using box and whisker plots. The breadth of each distribution represents the role of chance in manager selection, which has been highlighted by Molyboga and L'Ahelec (2016).

Exhibit 7 shows the distributions of the Sharpe ratios of the hypothetical portfolios, generated within the large-scale simulation framework for the out-of-sample period of January 2006–December 2017. The plot shows the same pattern: The MHRP approach improves out-of-sample performance, on average, while reducing downside risk.

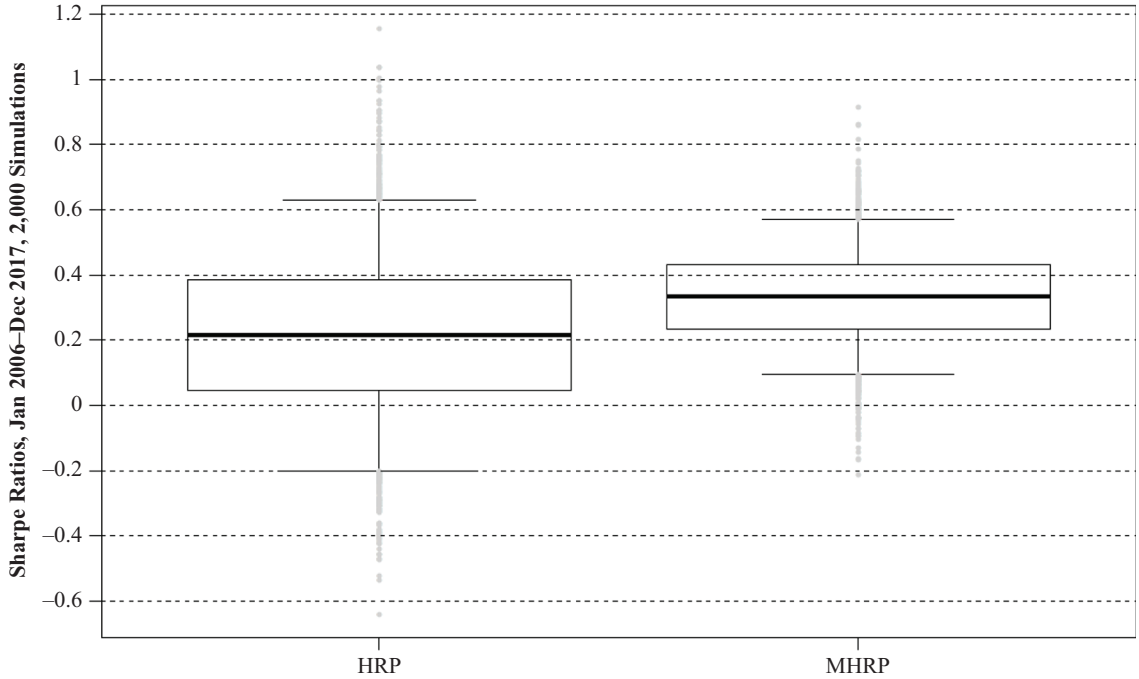
This finding is striking, yet intuitive, because the MHRP approach combines the structural benefits of the HRP approach of López de Prado (2016) and the practical ideas of improved covariance matrix estimation and diversification across time and portfolio constituents. The HRP method is a cutting-edge application of machine learning to portfolio management that overcomes the issues of unstable weights caused by an inversion of a near singular covariance matrix and a lack of alignment with the top-down portfolio management typically employed by sophisticated practitioners. The practical ideas tested here are simple yet commonplace among practitioners because of their robust performance benefits. As a result, the MHRP approach can be a potentially attractive portfolio management technique for any investor, provided its benefits are robust to real-life constraints.

Implications for Investors

In this section, I discuss the implications of using MHRP for investors in CTAs and other types of investments. The empirical results are relevant for CTA investors because the methodology of Molyboga and L'Ahelec (2016) accounts for several important constraints that institutional investors experience, such as size and length of track record, while seeking to produce implementable results by adjusting for data biases and eliminating the look-ahead bias that is present in most hedge fund studies. However, as discussed in detail by Molyboga (2019), the monthly rebalancing and monthly leverage changes that are required for implementing the MHRP approach are available for managed account investors but not for fund investors.[9]

[9] Fund investors cannot easily adjust portfolio leverage and face time gaps between a redemption from one fund and a subscription into another fund.

EXHIBIT 7
Out-of-Sample Sharpe Ratios of HRP and MHRP



Although the performance benefits of the MHRP approach should be carefully evaluated for other types of portfolios with their own sets of investments and real-life constraints, the three proposed enhancements are likely to have a positive impact on the portfolio performance of traditional and alternative investments because they improve the quality of the covariance matrix estimation and the diversification across time and portfolio constituents.

CONCLUDING REMARKS

In this article, I extend the HRP approach of López de Prado (2016) by adding three intuitive elements commonly used by practitioners. The proposed MHRP approach (1) replaces a sample covariance matrix with an exponentially weighted covariance matrix with Ledoit–Wolf shrinkage, (2) improves diversification across portfolio constituents within and across clusters by relying on an equal volatility allocation approach rather than an IVA approach, and (3) improves diversification across time by volatility targeting portfolios.

I find that MHRP improves the out-of-sample Sharpe ratios of multi-CTA portfolios by more than 50%, on average, while reducing downside risk. This performance improvement is relevant for institutional investors because it is implementable within a managed account investment and the empirical methodology thoroughly accounts for real-life constraints.

Future research can be dedicated to investigating the performance benefits of MHRP for portfolios of traditional and alternative investments and whether it can be further improved by employing alternative approaches to de-noising the correlation matrixes, such as the random matrix theory approach of Pafka, Potters, and Kondor (2004).

ACKNOWLEDGMENTS

I am grateful to Andrew Berkin, Adam Duncan, Frank Fabozzi, Christophe L'Ahelec, Marcos López de Prado, David Mascio, Sanida Mustafovic, Kevin Townsend, and Stephan Zeisberger for helpful comments.

REFERENCES

Ackermann, C., R. McEnally, and D. Ravenscraft. 1999. "The Performance of Hedge Funds: Risk, Return, and Incentives." *The Journal of Finance* 54 (3): 833–874.

Anson, M. 2011. "The Evolution of Equity Mandates in Institutional Portfolios." *The Journal of Portfolio Management* 37 (4): 127–137.

Bailey, D., and M. López de Prado. 2012. "Balanced Baskets: A New Approach to Trading and Hedging Risks." *Journal of Investment Strategies* 1 (4): 21–62.

Barroso, P., and P. Santa-Clara. 2015. "Momentum has its Moments." *Journal of Financial Economics* 116: 111–120.

Best, M. J., and R. R. Grauer. 1991. "On the Sensitivity of Mean-Variance-Efficient Portfolios to Changes in Asset Means: Some Analytical and Computational Results." *The Review of Financial Studies* 4: 315–342.

Bhardwaj, G., G. B. Gorton, and K. G. Rowenhorst. 2014. "Fooling Some of the People All of the Time: The Inefficient Performance and Persistence of Commodity Trading Advisors." *The Review of Financial Studies* 27: 3099–3132.

Bollerslev, T. 1986. "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics* 31 (3): 307–327.

Brinson, G. P., L. R. Hood, and G. L. Beebower. 1995. "Determinants of Portfolio Performance." *Financial Analysts Journal* 51 (1): 133–138.

Daniel, T., and T. J. Moskowitz. 2016. "Momentum Crashes." *Journal of Financial Economics* 122: 221–247.

DeMiquel, V., L. Garlappi, and R. Uppal. 2009. "Optimal Versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?" *The Review of Financial Studies* 22: 1915–1953.

Edelman, D., W. Fung, and D. A. Hsieh. 2013. "Exploring Unchartered Territories of the Hedge Fund Industry: Empirical Characteristics of Mega Hedge Fund Returns." *Journal of Financial Economics* 109: 734–758.

Foran, J., M. C. Hutchinson, D. F. McCarthy, and J. O'Brien. 2017. "Just a One Trick Pony? An Analysis of CTA Risk and Return." *The Journal of Alternative Investments* 20 (2): 8–26.

Hallerbach, W. 2012. "A Proof of Optimality of Volatility Weighting over Time." *Journal of Investment Strategies* 1: 87–99.

Joenvaara, J., R. Kosowski, and P. Tolonen. 2012. "New 'Stylized Facts' About Hedge Funds and Database Selection Bias." Working paper, University of Oulu, 2012.

Jorion, P., and C. Schwarz. 2017. "The Fix Is In: Properly Backing Out Backfill Bias." Working paper.

JPMorgan and Reuters. *RiskMetrics—Technical Document*, 4th ed. New York: JPMorgan. 1996.

Kim, A. Y., Y. Tse, and J. K. Wald. 2016. "Time Series Momentum and Volatility Scaling." *Journal of Financial Markets* 30: 103–124.

Ledoit, O., and L. Wolf. 2004. "Honey, I Shrunk the Sample Covariance Matrix." *The Journal of Portfolio Management* 30: 110–119.

López de Prado, M. 2016. "Building Diversified Portfolios That Outperform Out of Sample." *The Journal of Portfolio Management* 42 (4): 56–69.

MacKinlay, A. C., and L. Pastor. 2000. "Asset Pricing Models: Implications for Expected Returns and Portfolio Selection." *The Review of Financial Studies* 13: 883–916.

Maillard, S., T. Roncalli, and J. Teiletche 2010. "The Properties of Equally Weighted Risk Contribution Portfolios." *The Journal of Portfolio Management* 36: 60–70.

Markowitz, H. M. 1952. "Portfolio Selection." *The Journal of Finance* 7: 77–91.

Merton, R. C. 1980. "On Estimating the Expected Return on the Markets: An Exploratory Investigation." *Journal of Financial Economics* 8: 323–361.

Michaud, R. O. 1989. "The Markowitz Optimization Enigma: Is Optimized Optimal? *Financial Analysts Journal* 45: 31–42.

Molyboga, M. 2019. "Portfolio Management of Commodity Trading Advisors with Volatility Targeting." *Journal of Investment Strategies*, forthcoming.

Molyboga, M., S. Baek, and J. Bilson. 2017. "Assessing Hedge Fund Performance with Institutional Constraints: Evidence from CTA Funds." *Journal of Asset Management* 18: 547–565.

Molyboga, M., and C. L'Ahelec. 2016. "A Simulation-Based Methodology for Evaluating Hedge Fund Investments." *Journal of Asset Management* 17: 434–452.

Moreira, A., and T. Muir. 2017. "Volatility-Managed Portfolios." *The Journal of Finance* 72: 1611–1643.

Pafka, S., M. Potters, and I. Kondor. 2004. "Exponential Weighting and Random-Matrix-Theory-Based Filtering of Financial Covariance Matrices for Portfolio Optimization." http://arxiv.org/abs/cond-mat/0402573.

Pastor, L. 2000. "Portfolio Selection and Asset Pricing Models." *The Journal of Finance* 55: 179–223.

Yang, Z., K. Han, M. Molyboga, and G. Molyboga. 2016. "A New Diagnostic Approach to Evaluating the Stability of Optimal Portfolios." *The Journal of Investing* 25 (1): 37–45.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

The Evolution of Equity Mandates in Institutional Portfolios

Mark Anson
The Journal of Portfolio Management
https://jpm.pm-research.com/content/37/4/127

ABSTRACT: *The asset allocation decision is one of the most important determinants of overall fund performance for institutions around the world. Pension funds, endowments, and foundations spend considerable time deciding how to split up the trillions of dollars they collectively control. Inevitably, the largest allocation is to the equity slice of the pie. The reasons are several: higher returns, largest asset class in which to invest, good liquidity, easiest to understand, and the cheapest to transact in. Anson traces the evolution of equity mandates as institutional investors have become more sophisticated in their approach to the equity markets including international investing, global mandates, Fama–French risk factors, and risk parity.*

Just a One-Trick Pony? *An Analysis of CTA Risk and Return*

Jason Foran, Mark C. Hutchinson, David F. McCarthy, and John O'Brien
The Journal of Alternative Investments
https://jai.pm-research.com/content/20/2/8

ABSTRACT: *Recently, a range of alternative risk premium products has been developed, promising investors hedge fund/commodity trading advisor (CTA)-like returns with higher liquidity and transparency and relatively low fees. The attractiveness of these products rests on the assumption that they can deliver similar returns. Using a novel reporting bias–free sample of 3,419 CTA funds as a testing ground, the authors' results suggest that this assumption is questionable. They find that CTAs are not a homogenous group. They identify eight different CTA substrategies, each with very different sources of return and low correlation between substrategies. To illustrate the difficulty of modeling the strategies, they specify recently identified alternative risk premiums from the academic literature as factors to examine the sources of return of CTAs. They find that these premiums fail to explain between 56% and 86% of returns. Their results suggest that given the heterogeneity of CTAs, although these new products may deliver on liquidity, transparency, and fees, investors expecting hedge fund/CTA-like returns may be disappointed.*

Honey, I Shrunk the Sample Covariance Matrix

Olivier Ledoit and Michael Wolf
The Journal of Portfolio Management
https://jpm.pm-research.com/content/30/4/110

ABSTRACT: *The central message of this article is that no one should use the sample covariance matrix for portfolio optimization. It is subject to estimation error of the kind most likely to perturb a mean–variance optimizer. Instead, a matrix can be obtained from the sample covariance matrix through a transformation called shrinkage. This tends to pull the most extreme coefficients toward more central values, systematically reducing estimation error when it matters most. Statistically, the challenge is to know the optimal shrinkage intensity. Shrinkage reduces portfolio tracking error relative to a benchmark index and substantially raises the manager's realized information ratio.*

Submission Guidelines

PORTFOLIO MANAGEMENT RESEARCH

All submitted manuscripts must be original work that has not been published in another form such as another journal, magazine, or book chapter. An author may submit manuscripts for consideration by only one journal in the **Portfolio Management Research** (PMR) family at a time. Additionally, by submitting a manuscript for consideration by a PMR publication, an author agrees to refrain from submitting the manuscript elsewhere until an editorial decision has been made. If a manuscript is rejected, the author may submit it to one of the other journals in the PMR family.

PMR accepts submissions only through our online article submission system. To submit a manuscript, an author must upload it to the online system and complete the required checklist, including agreement to the terms of the PMR copyright agreement. We do not require any specific style or format for an initial submission. However, upon acceptance of an article for publication, the author will be expected to bring their manuscript into conformity with the stylistic and mechanical specifications described below. Also, at that time, the author will be expected to manually sign the PMR copyright agreement and to remove all prior versions of the article from everywhere that they have been posted, including SSRN and all other websites.

Target Length: The target length for articles is 4,000 words, not including the abstract, exhibits, and references. However, the editor of a journal may accept articles that are substantially longer or shorter. Articles shorter than 2,500 words or longer than 7,500 words are rarely accepted.

Title: You should give your article a short descriptive title. In general, an article's title should be no more than 12 words. You should try to avoid including words that carry little or no meaning.

Abstract: You should write a strong abstract for an article appearing in a PMR publication. The target length for abstracts is 160 words. The abstract should be non-technical, should not contain any references, and should define any acronyms on first use. An article's abstract is the only part of the article that appears on the unrestricted portion of the related journal's website. In the printed version of a journal, the abstract appears in the table of contents (as if it is written by the editorial team, not the authors). It does not appear directly before the article's main text.

Introductory Material: Please write strong introductory paragraphs. The introductory material is an important part of an article. It serves as a substitute for the abstract at the start of the article. Because an abstract is not presented right before its associated article, a reader will not necessarily have read the abstract before reading the main text. An article's introductory material should describe the article's main points and briefly explain the reasoning or analysis that supports them. The introductory material should explain why the findings, results, or conclusions are important and how they add to prior research in the area. If applicable, the introductory material should convey how an investment professional can use the findings, results, or conclusions in a practical way. In most cases it is helpful to include a "roadmap" paragraph at the end of the introductory material. The purpose of the roadmap paragraph is to inform readers about how the article is organized into different sections and what is covered in each one.

Headings: Articles in PMR publications can have up to three levels of headings. The first heading in an article should appear after the introductory material. Articles in PMR publications do not include any heading before an article's introductory material.

Conclusion: Please end the main text of an article with a short conclusion that briefly summarizes the article's main point. The conclusion may resemble the article's abstract, though it need not omit references and technical terms used in the article. The conclusion should not introduce new material.

Footnotes: PMR publications now use footnotes rather than endnotes. Footnotes generally should not be used for references but rather for peripheral or tangential commentary that would not be suitable for inclusion in the main text.

References: PMR follows the "author-date" reference format of *The Chicago Manual of Style* (CMS), 17th Edition, Chapter 15 both for references included in the text of an article and for the list of references at the end of an article. The list of references should immediately follow the main text.

Style, Usage, and Grammar: PMR follows CMS with respect to matters of style, usage, and grammar. This includes standards for hyphenation, abbreviations, and capitalization.

Exhibits: Please call all tabular or graphical material Exhibits. Number exhibits with Arabic numbers consecutively in order of appearance in the text. Do not call them Table 1 and Figure 1. An article's text should refer to all of its exhibits. In general, the text should clearly explain the point supported or demonstrated by each exhibit. Additionally, except for the simplest tables and charts, the text should explain how to read each exhibit. For tables that report statistical results, the text should direct readers to the key values and clearly explain what they mean. For an equation, the text should describe and explain the relationship represented by the equation and clearly define all terms used in the equation. In many cases, it is better to place content that is heavy with equations in an appendix. Within the PMR family of journals, *The Journal of Fixed Income* and *The Journal of Derivatives* generally allow for the highest proportions of quantitative material in the main text.

Exhibit Presentation: Authors are encouraged to submit articles with exhibits in color. Although PMR publications are generally printed in black and white, the online version can show color exhibits. Also, upon article acceptance, if you wish to have your exhibits printed in color, you may do so at a cost of $1,200. Please make sure that all categories in an exhibit can be distinguished from each other and match the categories in any legend connected with an exhibit. If you use color, please make sure that each exhibit will be legible if printed in black-and-white. This is important for all articles because a reader may print an article on a black-and-white printer.

Copyright Agreement: PMR's copyright agreement form must be signed prior to publication. Only one author's signature is necessary. Should your article not be accepted the copyright agreement will be considered null and void. Upon acceptance of the article, no further changes are allowed, except with the permission of the editor.

the journal of PORTFOLIO management

NOVELLUS PERICULA

Novel Risks & Volatility: Identification and Measurement Challenges for Portfolio Management

special issue—call for papers

Although some novel risks, such as climate risk, cybersecurity risk, and geopolitical risk, may be labeled as non-financial, their impact needs to be measured financially. The recent global pandemic highlights the challenges in identifying emerging risks and measuring their impact on volatility. This special issue of *The Journal of Portfolio Management* will include impactful research focusing on the identification and measurement of novel risks with specific attention to their implication for portfolio management.

editor
FRANK J. FABOZZI
Professor of Finance, EDHEC Business School

guest co-editor
AHMET K. KARAGOZOGLU
C.V. Starr Distinguished Professor of Finance, Hofstra University
Visiting Scholar, Volatility and Risk Institute, New York University

Find out more and submit your research at
jpm.pm-research.com/novellus-pericula

Gain traction in the field of sustainable investing

THE JOURNAL OF
impact&esg
INVESTING

launching September 2020

With environmental, social, and governance factors becoming increasingly important when making investment decisions and managing risk, our new journal will ensure you have all the research you need to manage and enhance your portfolios effectively.

PORTFOLIO
MANAGEMENT
RESEARCH

JESG.pm-research.com